Vol. XXXV. Parts I and II May, 1948

BIOMETRIKA

FOUNDED BY
W. F. R. WELDON, FRANCIS GALTON AND KARL PEARSON

MANAGING EDITOR
E. S. PEARSON

ASSOCIATE EDITORS
M. G. KENDALL JOHN WISHART

in consultation with
HARALD CRAMÉR J. B. S. HALDANE
R. C. GEARY G. M. MORANT
MAJOR GREENWOOD

ISSUED BY
THE BIOMETRIKA OFFICE, UNIVERSITY COLLEGE, LONDON
PRINTED AT THE UNIVERSITY PRESS, CAMBRIDGE

Reprinted by offset-litho 1953, 1963

[*Issued* 10 *June* 1948]

UNIVERSITY OF UTAH LIBRARIES

JAMES FOWLER TOCHER
1864–1945

VOLUME XXXV, PARTS I AND II MAY 1948

JAMES FOWLER TOCHER

James Fowler Tocher (1864–1945), chemist, ethnologist, biometrician, agriculturist and man of affairs, began adult life by opening a chemist's business in Peterhead more than sixty years ago. Even sixty years ago, scientific chemistry was not restricted to professors or even analysts, and certainly proprietors of businesses sometimes made money. The Southron, however, unduly influenced by jokes about Aberdonians—most of which are based on the principle, *lucus a non lucendo*—would not think Peterhead a promising venue either for scientific research or earning a competence. In fact, the business was a financial success; Tocher disposed of it in 1912 and when he passed to the analytical branch of professional chemistry, he had already made a name among chemists and biometricians (he was President of the Pharmaceutical Society of Great Britain in 1908).

Like several other men who have done important scientific work Tocher owed a good deal to a local society, the Buchan Field Club. In the south, although the name Field Club is used, e.g. the Essex Field Club, Natural History Society is perhaps a commoner designation and can include all the descriptive sciences. It is quite possible that these societies are educationally almost as valuable as classes in technical colleges even now; sixty years ago they were invaluable. The Buchan Field Club published transactions; in these appeared Tocher's earliest statistical papers, the very first in 1895.

They were concerned with the ethnology of Buchan and form the basis of the wider studies with which his name is usually associated. Among the earlier papers is one, written jointly with a frequent collaborator, Mr, afterwards Prof., James Gray, on the 'Frequency and Pigmentation Value of Surnames in East Aberdeenshire'.

More than ninety years ago, William Farr wrote an essay on the statistics of surnames which appeared in the *Sixteenth Annual Report of the Registrar General* and may have been read by Tocher because it is reprinted in the memorial volume edited by Noel Humphreys which preserves some, but by no means all, of the delightful essays by an old master. Farr used the registers of births and deaths and found in a sample of 275,405 names 32,818 different surnames or 11·9 % different surnames. In Wales the proportion was much smaller. 'The name of John Jones is a perpetual incognito in Wales, and being proclaimed at the cross of a market town would indicate no one in particular.' Farr touched on the local distributions and asked such questions as whether the present predominance of the Smiths were due to the original numerical strength of that great family, or 'to some special circumstances acting upon the ordinary laws of increase, owing to which the descendants of the hammer-men have multiplied at a greater rate than the bearers of any other name?' Did the progeny, he asked, of the tawny Browns increase faster than that of the fair-complexioned Whites?

Tocher interested himself in such problems. The data of the investigation of Gray & Tocher (printed in 1902) were pigmentation records of 14,561 school-children in East Aberdeenshire. These 14,561 children had 751 surnames, 5·2 %. A smaller proportion than in Farr's data, as one might expect, since in a school population there would be many more brothers and sisters than in the birth and death registers for two quarters. In East Aberdeenshire, Smith, scoring 203, was easily beaten by Milne with 267; third came Taylor, which is fourth on Farr's list for England and Wales. Gray & Tocher worked out the mean and standard deviation in pigmentation units of the children with different surnames. 'The

blondest surname in our list is Pirie....Next to Pirie comes the surname Wallace. This points to the conclusion that the Wallace sept sprang from ancestors with a decided blonde tendency.'

The use of the standard deviation in this paper, as well as an allusion to Karl Pearson's work on assortative mating, show that either Gray or Tocher or both were already readers of K.P. At what date Tocher made the personal acquaintance with Karl Pearson, which ripened into a warm friendship, I do not know. Five years later, in 1907,* Tocher's first full-scale biometric memoir appeared in this *Journal* (Vol. **5**, pp. 298–350). The original data are printed as an appendix to the volume. It is careful biometric study of the stature, pigmentation, and craniometry of inmates of Mental Homes in Scotland. In 1908 (Vol. **6**, pp. 129–235, Appendix 1–63) is an equally careful study of the pigmentation of school-children in Scotland. These researches were financed by the Henderson Trust who published in book form (Oliver & Boyd, 1924) a summary by Tocher of the results recorded in the earlier papers together with those of later anthropometric observations on samples of the civil populations of Aberdeenshire, Banffshire and Kincardineshire and on soldiers of Scottish nationality.

The use of superlatives is not a method of science—there is *no* greatest 'scientist'—but, whether any anthropometric survey of another part of the British Isles so far undertaken can take precedence over the voluntary work of Tocher supported financially not by the state but by a private trust, is a question which may fairly be asked. I think the answer is 'no'. Tocher's analysis followed the lines of teaching so many of us older people look back upon with gratitude. It may well be that, since his time, the technique of craniometry has been extended and genetic research has modified some of the perhaps rather naïve theories of inheritance popular a generation ago. But Tocher's work remains of fundamental importance; it is a pity his example did not inspire Englishmen to extend his study of mental hospital populations.

In 1925 (*Biometrika*, Vol. **17**, pp. 142–58) R. Greenwood, C. M. Thompson & H. M. Woods published a memoir on heights and weights of patients and there are scattered through the literature other statistical papers, but not, I think, any large-scale investigation. But the mind-body relation still interests us all. Our remote predecessors thought there was a corporeal basis of the sanguine, melancholic, choleric and phlegmatic temperaments; psychologists and pathologists of to-day reject the bases imagined by our ancestors, but still search for *a* basis. Most people know that pulmonary tuberculosis takes a heavy toll in mental hospitals and perhaps dismiss the subject with some vague remark about the 'unfit' or, alternatively, remember that the old physicians thought that grief or emotional shock was a factor predisposing to consumption. There seems to be no doubt that a particular form of mental disorder, dementia praecox, is particularly associated with tuberculosis, so intimately that one writer has maintained that dementia praecox and pulmonary tuberculosis can, in a sense, be regarded as modifications or expressions of a common defect or diathesis. All this is, however, the merest speculation without a biometric study. Obviously to take the heights and weights of patients in institutions would only be a beginning. Tocher did not classify his data under diagnosis, perhaps he could not have done so usefully because,

* [During the summer of 1906, Tocher came south to Danby, in the Yorkshire moors, where K.P. was spending his vacation and the final arrangements for the publication of this memoir were no doubt then discussed. Tocher was also a friend of W. R. Macdonell whom he succeeded, at a rather later date, as part time lecturer in Statistics at the University of Aberdeen. *Ed.*]

although 4436 males and 3951 females sound large numbers, when they are subdivided into age- and diagnosis-groups, the classes would grow small. The records of the mental hospitals of Great Britain in the forty years since Tocher did his work would suffice. But there has not been an English Tocher.

In his later years, Tocher's deserved reputation led to official or semi-official appeals for his help in work involving statistical analysis. He was very active in the study of milk production; in tackling such problems his chemical training was of great value. A matter of obvious importance is whether poor yield is due to nature or nurture or to both. Tocher was of opinion that both factors were involved but that 'the solution to the problem of deficient milk must be the careful selection of cows with good milk records and a good milking pedigree' (see Stewart & Tocher, *J. Dairy Res.*, January 1936). Tocher did not forget, as some academic writers did, the importance of vicious circles; those who farm with insufficient capital cannot afford to buy pedigree stock and cannot afford to feed adequately what stock they have. He protested vigorously against the gross injustice of any legal presumption that because a particular sample of milk had a percentage of butter fat or a percentage of solids-not-fat below a prescribed minimum, the milk had been adulterated. It might, he thought, be quite reasonable to forbid the marketing of such milk, but certainly not right to affix a stigma to the farmer. Tocher was a good biometrician and a good practical Christian.

In attempting to give a personal impression of Tocher, a writer who also knew two of his Aberdonian friends, Charles Creighton (1847–1927) and William Bulloch (1868–1941), both of whom left classical contributions to scientific literature, is tempted to compare them. Creighton was, in the old-fashioned sense of the word, the greatest scholar of the three. His *History of Epidemics in Britain* is a classic and other of his less-known writings would have been approved by Dr Johnson. Bulloch, less familiar than Creighton with the ancient writers, had an encyclopaedic knowledge of modern pathological literature and readers of this *Journal* are likely to remember his contribution to the *Treasury of Human Inheritance*, his study of Haemophilia which, incidentally, contains many illustrations of the irony which made Bulloch a famous raconteur. I do not think Tocher was the equal of these two friends in scholarship or literary power, but—although to some who perhaps over-value formal scientific training it may sound paradoxical—I should say he was a better, a more original, *scientific* investigator than either. Epidemiologists will never forget Creighton, bacteriologists will never forget Bulloch. Biometricians, I hope, will read Tocher's papers and show their gratitude in the way he would most have appreciated, that is by completing some of the tasks to which he put his hand.

Tocher had always been a good friend of this *Journal* and he became a Trustee on the inception of the Biometrika Trust in 1936.

MAJOR GREENWOOD

BIBLIOGRAPHY OF TOCHER'S PAPERS

1889. The springs and wells of Peterhead. *Trans. Buchan Fld Cl.* **1**.
1891. Test for the detection of sesame oil in olive oil and isolation of another substance (sesamin) from sesame oil. *Pharm. J.* **21** (24 Jan.).
1893. A further note on sesamin. *Pharm. J.* **23** (25 Feb.).
1894. The production of gas from paraffin oils, and from pure members of paraffin and terpene series. *J. Soc. chem. Ind., Lond.*, **13** (31 Mar. and 31 May).
1895. (With John Gray.) The ethnology of Buchan. Preliminary ethnographical observations. *Trans. Buchan Fld Cl.* **3**.

1897*a*. The ethnology of Buchan. Part II. Ethnographical survey of school-children in Buchan. *Trans. Buchan Fld Cl.* **4**.
1897*b*. Report on scales of contributions and systems of levies. British Order of Ancient Free Gardeners.
1900*a*. (With John Gray.) The physical characteristics of adults and school-children in East Aberdeenshire. *J. R. anthrop. Inst.* **30** (=N.S. **3**).
1900*b*. (With John Gray.) The physical characteristics of the population of West Aberdeenshire. *J. R. anthrop. Inst.* (Reviews and Miscellanea), **30**, no. 84 (=N.S. **3**).
1900*c*. The volumetric determination of red lead. *Pharm. J.* **64** (24 Mar.).
1900*d*. Address on the Free Gardeners and the Friendly Society Movement.
1910*a*. (With John Gray.) The ethnology of Buchan. Part III. The physical characteristics of adults and school-children in East Aberdeenshire. *Trans. Buchan Fld Cl.* **6**. (Contains the same data as 1900*a*, with some additions.)
1901*b*. The volumetric determination of phenol. *Pharm. J.* **66** (23 Mar.).
1902*a*. (With John Gray.) The ethnology of Buchan. Part IV. The frequency and pigmentation value of surnames in East Aberdeenshire. *Trans. Buchan Fld Cl.* **7**.
1902*b*. (With John Gray.) The physical characteristics of the eskimo of Southampton Island, etc. *Trans. Buchan Fld Cl.* **7**.
1902*c*. Scottish universities and pharmaceutical education. *Pharm. J.* **68** (1 Mar.).
1902*d*. The oxidation and determination of uric acid and urates. *Pharm. J.* **69** (16 Aug.).
1903*a*. The characteristics of mankind. *Aberdeen Free Press* (21 Mar.).
1903*b*. A scheme of representation by territorial districts on the Pharmaceutical Council. *Pharm. J.* **71** (15 and 22 Aug.).
1905*a*. Medical inspection of schools. Address to the 31st Annual Congress of the Incorporated Sanitary Assoc. of Scotland. Inverness, Sept. 1905.
1905*b*. Anthropometric survey of the inmates of asylums in Scotland. *Henderson Tr. Rep.* **1**.
1906*a*. The activity of pepsin after brief contact with certain inorganic compounds. *Pharm. J.* **77** (28 July).
1906*b*. The detection of citrates and tartrates. *Pharm. J.* **77** (28 July).
1907*a*. The anthropometric characteristics of the inmates of asylums in Scotland. *Biometrika*, **5**, 298–350 and Suppl. (Supplement is a reprint of 1905*b* above.)
1907*b*. Anthropometry, its aims and methods. Privately printed.
1908. Pigmentation survey of school-children in Scotland. *Biometrika*, **6**, 129–235 and Suppl.
1909. Some problems of interest to pharmacists to-day. *Year Book of Pharmacy and Trans. of the British Pharmaceutical Conference.*
1910. The necessity for a national eugenic survey. *Eugen. Rev.* **2**.
1913. The criteria for purity of water—chemical and bacteriological. *Proc. Aberdeen Ass. Civil Engrs*, **13** (1912–13).
1915. (With Karl Pearson.) On criteria for the existence of differential death-rates. *Biometrika*, **11**, 159–84.
1918. Food values and costs in relation to milk production. *Scot. J. Agric.* **1**, no. 1.
1919*a*. Investigation into the milk yield of Ayrshire cows. *Trans. Highl. agric. Soc. Scot.*
1919*b*. Variations in the composition of milk. *Scot. J. Agric.* **2**, no. 3.
1919*c*. Science and the struggle for existence. *The Rotary Wheel.*
1919*d*. Report on distillery by-products. Privately printed.
1921. On the need for consolidating the law relating to the sale of food and drugs. *Trans. inc. sanit. Ass. Scot.*
1922*a*. The relationship between citric solubility of phosphates and yield of turnip crop. *Trans. Highl. agric. Soc. Scot.*
1922*b*. The citric solubility of mineral phosphates. *J. agric. Sci.* **12**, pt. 2.
1923*a*. Grass sickness in horses. (Reports by J. F. Tocher, J. W. Tocher, W. Brown and J. B. Buxton.) *Vet. Rec.* **35**.
1923*b*. Milk yields and associated factors. *Proc. World Dairy Congr.*
1924*a*. Grass sickness in horses. *Trans. Highl. agric. Soc. Scot.* (Contains same data as 1923*a* above.)
1924*b*. Anthropometric observations on samples of the civil population of Aberdeenshire, etc. *Henderson Tr. Rep.* **2**.
1924*c*. A study of the chief physical characters of soldiers of Scottish nationality and a comparison with the physical characters of the insane population of Scotland. *Henderson Tr. Rep.* **3**.
1925*a*. Variations in the composition of milk. *Edin. Med. J.* (May).
1925*b*. Variations in the composition of milk. H.M.S.O.
1926*a*. The pharmacist, his training and vocation. *Pharm. J.* **117** (9 Oct.).

1926*b*. Errors of judgment in chemical analysis. *Analyst* (July).
1926*c*. Variations in the composition of milk. *Analyst* (Dec.).
1927*a*. Causes of variation in the proportion of butter fat in milk. *Scot. J. Agric.* **10**, no. 1.
1927*b*. Variation in the proportion of solids-not-fat in milk. *Scot. J. Agric.* **10**, no. 2.
1927*c*. Sugar in milk. *Scot. J. Agric.* **10**, no. 4.
1928. An investigation of the milk yield of dairy cows. *Biometrika*, **20**B, 105–244.
1929. The sale of milk. Suggested amendments of the law. *Scot. J. Agric.* **12**, no. 4.
1930. The study of literature. Lecture to Peterhead Literary Soc.
1931*a*. Should the law relating to the sale of milk be amended? *Trans. Yorks. agric. Soc.*
1931*b*. What is probable error? Lecture to Institute of Chemistry of Great Britain and Ireland.
1932. A New Year's Message. *Chem. & Drugg.* (2 Jan.).
1934*a*. Raw versus pasteurized milk. *Scot. J. Agric.* **17**, no. 1.
1934*b*. The services of Francis Galton and his school to physical anthropology and eugenics. *Rep. Brit. Ass.*, Section H.
1935. The proportions of certain poisonous substances in feeding stuffs and their effect on livestock. *Vet. Rec.* **47**.
1936. (With Alice Stewart.) The effect of variations in feeding on dairy cows yielding milk of poor quality. *J. Dairy Res.* **7**, no. 1.
1937. Agricultural education and research in Scotland. *Aberdeen Univ. Rev.* (Mar.).
1939*a*. Pasteurization of milk. *Trans. Yorks. agric. Soc.* no. 96.
1939*b*. Mortality among livestock due to poisons. *Vet. Rec.* **51**.
1940. The role of science. *The Second Book of Buchan.*
1941*a*. Prehistory and early history. *The Second Book of Buchan.*
1941*b*. The bracken problem. *Trans. Highl. agric. Soc. Scot.*

ON SOME MODES OF POPULATION GROWTH LEADING TO R. A. FISHER'S LOGARITHMIC SERIES DISTRIBUTION

By DAVID G. KENDALL, M.A.
Magdalen College, Oxford

1. R. A. Fisher (1943) in a co-operative study written with A. S. Corbet and C. B. Williams has developed a mathematical theory which describes with some success the relative numbers of animals of different species obtained when sampling at random from a heterogeneous population. This problem was first considered in relation to (i) Corbet's work on the distribution of butterflies in the Malay Peninsula, and (ii) the numbers of moths of different species caught in a light-trap over a given period of time (Williams's data). Fisher began by assuming that for a particular species the number of individuals caught in time t would be distributed as a Poisson variable of expectation ωt, where ω may be called the *intrinsic abundance* of the species. He suggested that ω might be distributed in the Eulerian (or χ^2) form

$$\frac{1}{\Gamma(k)}\left(\frac{k}{\Omega}\right)^k e^{-k\omega/\Omega}\omega^{k-1}\,d\omega \quad (0<\omega<\infty), \tag{1}$$

where Ω is the mean value of ω and k is a constant parameter, and showed that the actual number caught would then follow a negative binomial distribution with index k*. In fitting such a distribution to Corbet's data he obtained very small values of k, and this suggested that it might be worth while examining what would happen if Ω and k were allowed to tend to zero in a constant ratio. In this way Fisher found that if a species were known to have been caught, it would be represented in the catch by exactly n individuals with a probability

$$\frac{x^n}{ny} \quad (n = 1, 2, 3, \ldots), \tag{2}$$

where† $$y = \ln\frac{1}{1-x}, \quad x = \frac{at}{1+at} \quad \text{and} \quad a = \lim \Omega/k.$$

The success of this 'logarithmic series distribution' in graduating the entomological data of Corbet, Williams and others implies that in the populations concerned the distribution of intrinsic abundance must be (for ω not too small) effectively of the form

$$Ae^{-\omega/a}\,d\omega/\omega \quad (A = \text{constant}). \tag{3}$$

(This cannot of course be true for *all* ω, for then the integral of total probability would not converge.) It will be noticed that the distribution (3) of intrinsic abundance is itself the continuous analogue of the logarithmic series. The success of (2) in describing the relative numbers of individuals caught is thus a challenge to biologists to provide a theoretical interpretation for (3).

In this connexion it is worth noting that if one is concerned with a population containing only a finite number (Z, say) of species, then the continuous distribution (3) can be replaced by a logarithmic series, and results similar to those of Fisher follow as before. Thus, suppose that the actual number ν of individuals by which a particular species is represented *in the whole population* is distributed in the discrete form

$$\frac{X^\nu}{\nu Y} \quad (\nu = 1, 2, 3, \ldots), \tag{4}$$

* This step in the argument is of course equivalent to that taken by M. Greenwood & G. U. Yule (1920) in another context.

† I write $\ln z$ for the natural logarithm of z.

where $Y = -\ln(1-X)$, and let $p = 1-e^{-\gamma t}$ be the chance that an individual will be caught in an exposure of duration t. Then the chance that the species will have $n = 0, 1, 2, 3, \ldots$ representatives *in the sample* is given by*

$$P_0 = 1 - \frac{y}{Y} \quad \text{and} \quad P_n = (1-P_0)\frac{x^n}{ny} \quad (n>0), \tag{5}$$

i.e. a logarithmic distribution with a zero term added. Here

$$x = 1-e^{-y} \quad \text{and} \quad e^y - 1 = (e^Y - 1)(1-e^{-\gamma t}), \tag{6}$$

and so $y \leqslant Y$ for all t; for very small exposures p will be small and then

$$y \simeq \gamma t(e^Y - 1),$$

while for very long exposures p will be nearly equal to unity and then

$$y \simeq Y.$$

The expected values of S and N (the number of species and the number of individuals in the catch) are

$$\bar{S} = Z(1-P_0) = yZ/Y \quad \text{and} \quad \bar{N} = (e^y - 1)Z/Y.$$

Thus as $t \to \infty$, $\bar{S} \to Z$ and $\bar{N} \to (e^Y - 1)Z/Y$, while for *all* values of t

$$\bar{S} = \alpha \ln(1 + \bar{N}/\alpha), \tag{7}$$

where $\alpha = Z/Y$ is a constant independent of the time of exposure and corresponding to Fisher's 'index of diversity'. Formula (7) is, in fact, identical with the well-known result due to Fisher (1943), although the derivation given here proceeds from somewhat different assumptions.

Williams (1944*a*, *b*) has shown that the logarithmic series (2) can also be applied to a great variety of other biological problems, in which the integer n is variously the number of species per genus, the number of genera per subfamily, the number of parasites per host, and even the number of research papers per biologist (published in a particular year). It is hard to believe that a single mechanism will be found to explain the relevance of the logarithmic series to *all* these problems, and it seems therefore well worth while to record any theoretical models which may be found to lead mathematically to this distribution. In the remainder of this paper I shall describe a number of discontinuous Markoff processes which lead to distributions of negative binomial and logarithmic series form, in the hope that some of these may be found to be of biological significance.

2. The stochastic processes to be considered here will for convenience be described in relation to the growth of a hypothetical population of organisms, whose numbers fluctuate with the incidence of mortality, reproduction (by binary fission) and immigration from the outside world. Let n be the size of the population at time t; then $n(t)$ is a random function which develops in the following manner:

(i) If $n > 0$, the only possible transitions in an element of time dt are from n to $n-1$, n or $n+1$, and the transition probabilities are

$$n \to \begin{cases} n+1 & \text{with probability} \quad (n\beta + \kappa)\,dt, \\ n & \text{with probability} \quad 1-(n\mu + n\beta + \kappa)\,dt, \\ n-1 & \text{with probability} \quad n\mu\,dt. \end{cases}$$

* It appears that this sampling property of the logarithmic series distribution (which is easily proved with the aid of the generating function) has already been noticed by C. B. Williams (1947) and M. H. Quenouille.

(ii) (This is actually included in (i), but an explicit statement is desirable.) If $n = 0$, the only possible transitions in time dt are from 0 to 0 or 1, the transition probabilities being

$$0 \to \begin{cases} 1 \text{ with probability } & \kappa dt, \\ 0 \text{ with probability } & 1-\kappa dt. \end{cases}$$

Let $P_n(t)$ be the probability that at time t the population size is n; then it is possible to set up an infinite system of differential-difference equations which together with the distribution $\{P_n(t_0)\}$ at some initial time $t = t_0$ determine the $\{P_n(t)\}$ at all subsequent times, and so govern the mode of growth of the population. Two alternative sets of initial conditions will be considered here:

(A) $$P_0(-T) = 1 \quad \text{and} \quad P_n(-T) = 0 \quad (n > 0).$$

This implies that at time $t = -T$ the population size was zero.

(B) $$P_0(-T) = 0, \quad P_1(-T) = 1 \quad \text{and} \quad P_n(-T) = 0 \quad (n > 1).$$

This implies that the population commenced with one individual at time $t = -T$.

Next it is necessary to give a biological interpretation of the effects associated with the constants β, μ and κ.

The first of these, β, represents the reproductive power of the individuals composing the population, the effects of sex and age being ignored. Thus it is supposed that if attention is focused on any one individual at time t, it will be found to undergo binary fission at a time $t+\tau$, where τ has the probability distribution

$$e^{-\beta\tau}\beta d\tau \quad (0 < \tau < \infty). \tag{8}$$

An important consequence of the assumption (8) is that the time to the next subdivision, for any individual, is statistically independent of its past history, and in particular it is independent of the length of time since that individual was itself formed by the fission of its parent. At first sight it might appear that a bacterial colony would provide a good example of such a population growing by binary fission, but it must be remembered that the generation times of bacteria, while liable to considerable random variation, have a frequency distribution* very different from (8) and possessing a pronounced non-zero mode.

The n individuals present at any time are assumed to reproduce themselves independently of one another, and at the same constant mean rate. At each subdivision the parent can be thought of either as being replaced by its two offspring, or as only adding one new member to the colony and remaining a member itself; a transition $n \to n+1$ then takes place.

In a similar way the constant μ represents the loss to the colony due to 'mortality'. It is assumed that an individual does not lose its power to reproduce unless it 'dies', and that it then ceases to be regarded as a member of the colony, so that a transition $n \to n-1$ takes place. Such a transition could, however, also mean the removal (by any means) of an individual from the region considered; these two sources of loss are mathematically indistinguishable and will therefore be covered by the same symbol μ. Thus if an individual is observed at time t, it will disappear from the population at a time $t+\tau$, where τ has the distribution

$$e^{-\mu\tau}\mu d\tau \quad (0 < \tau < \infty).$$

* See, for example, Kelly & Rahn (1932) and Hinshelwood (1946).

The β- and μ-effects, described separately, are to be thought of as acting simultaneously and independently one of the other. Thus, when the β- and μ-effects are acting together, the chance that an individual remains inactive for a time τ and then subdivides during the subsequent time interval $d\tau$ is

$$e^{-(\beta+\mu)\tau}\beta\,d\tau.$$

Integration from $\tau = 0$ to $\tau = \infty$ then gives the chance that the individual will subdivide before the μ-effect has removed it from the colony; this is

$$\int_0^\infty e^{-(\beta+\mu)\tau}\beta\,d\tau = \frac{\beta}{\beta+\mu}.$$

Finally the κ-effect is one of 'immigration from outside', i.e. it is supposed that from time to time individuals not initially members of the colony may join it and proceed to behave exactly like the other members. If, from a given time instant t, the next such 'immigration' occurs at time $t+\tau$, the distribution of τ is assumed to be

$$e^{-\kappa\tau}\kappa\,d\tau \quad (0<\tau<\infty).$$

The structure of the model will now be clear. It only remains to point out that the probabilities of a positive unit increment from the β-effect or a negative unit increment from the μ-effect in an element of time dt will each be proportional to n, the existing population size, while the chance of a unit positive increment from the κ-effect is the same for all values of n.

3. The differential-difference equations of the process can now be written down. They are

$$\frac{d}{dt}P_n(t) = (n+1)\mu P_{n+1}(t) - \{n(\beta+\mu)+\kappa\}P_n(t) + \{(n-1)\beta+\kappa\}P_{n-1}(t), \tag{9}$$

if $n \geqslant 1$, and

$$\frac{d}{dt}P_0(t) = \mu P_1(t) - \kappa P_0(t). \tag{10}$$

It is convenient to *define* $P_n(t)$ as being identically equal to zero when $n<0$; equation (10) can then be included within the general form (9). I owe to Dr M. S. Bartlett the remark that systems of equations of this type can most conveniently be solved with the aid of the generating function

$$\phi(z,t) \equiv \sum_{n=-\infty}^{\infty} z^n P_n(t). \tag{11}$$

It will be seen from (9) that $\phi(z,t)$ must satisfy the partial differential equation

$$\frac{\partial\phi}{\partial t} = \{\mu-(\beta+\mu)z+\beta z^2\}\frac{\partial\phi}{\partial z} + \kappa(z-1)\phi, \tag{12}$$

which together with one of the boundary conditions,

(A) $$\phi(z,-T) = 1,$$

or

(B) $$\phi(z,-T) = z,$$

and the requirement that the expansion of ϕ must contain no terms in $1/z, 1/z^2, \ldots$, is sufficient completely to determine the process.

The differential equation (12) is of the standard Lagrangian form, the auxiliary equations being

$$-dt = \frac{dz}{(\beta z-\mu)(z-1)} = -\frac{d\phi}{\kappa(z-1)\phi}. \tag{13}$$

'First integrals' are $(\beta-\mu)t+\ln(z-1)-\ln(\beta z-\mu)=\text{constant}$,

and $\kappa\ln(\beta z-\mu)+\beta\ln\phi=\text{constant}$,

if $\kappa>0$ and $\beta\neq\mu$, and so the general integral of (12) is then

$$\phi(z,t)=(\mu-\beta z)^{-\kappa/\beta}\Phi\left\{\frac{\mu-\beta z}{1-z}e^{-(\beta-\mu)t}\right\}, \tag{14}$$

where Φ is an arbitrary function to be determined from the boundary conditions.

With boundary condition (A) it will be found that

$$\phi(z,0)=\left(\frac{\beta-\mu}{\beta\Lambda-\mu}\right)^{\kappa/\beta}\left\{1-\frac{\beta(\Lambda-1)}{\beta\Lambda-\mu}z\right\}^{-\kappa/\beta}, \tag{15}$$

where Λ has been written for $e^{(\beta-\mu)T}$ (this is equal to the *expected* factor by which the population will be multiplied in a time interval T, when it is growing in the absence of immigration). Similarly with boundary condition (B) one obtains

$$\phi(z,0)=\frac{(\beta-\mu)^{\kappa/\beta}}{(\beta\Lambda-\mu)^{1+\kappa/\beta}}\{\mu(\Lambda-1)-(\mu\Lambda-\beta)z\}\left\{1-\frac{\beta(\Lambda-1)}{\beta\Lambda-\mu}z\right\}^{-1-\kappa/\beta}, \tag{16}$$

where Λ has the same meaning as before.

When $\kappa=0$, the solutions are of a slightly different form. The general integral of (12) is then

$$\phi(z,t)=\Phi\left\{\frac{\mu-\beta z}{1-z}e^{-(\beta-\mu)t}\right\}. \tag{17}$$

Condition (A), as might be expected, gives $\phi(z,t)\equiv1$; this merely asserts that a zero population will remain zero if there is no immigration. Condition (B), however, gives

$$\phi(z,0)=\frac{\mu(\Lambda-1)-(\mu\Lambda-\beta)z}{\beta\Lambda-\mu}\left\{1-\frac{\beta(\Lambda-1)}{\beta\Lambda-\mu}z\right\}^{-1} \tag{18}$$

It will be noticed that in every case the solution is a regular function of z near $z=0$, so that in the Laurent expansion the coefficients of the negative powers will all vanish, as required.

4. It is now a simple matter to interpret these solutions. There are three cases of special interest.

(i) Consider first a population growing from zero, so that (A) is the appropriate boundary condition, the population being established in the first instance by immigration from outside. The κ-effect is of course acting all the time, even after the colony has started growing, so that in general there will at any time be present a number of independent families, each descended from a different immigrant ancestor. According to (15) the population size n, after the process has been developing for a time T, will be distributed as a negative binomial variate with index κ/β and mean value

$$E(n)=\frac{\kappa(\Lambda-1)}{\beta-\mu}. \tag{19}$$

Thus if $\beta>\mu$, the expected size of the population will for large T grow geometrically at an exponential rate $(\beta-\mu)$.

If, however, $\beta<\mu$, so that the force of mortality more than compensates for the force of reproduction, one will have

$$\lim_{T\to\infty}E(n)=\frac{\kappa}{\mu-\beta}. \tag{20}$$

This is the mean of the stable distribution of population size which can just be maintained by the immigration rate κ. If κ were equal to zero the population would almost certainly die out in a finite time.

(ii) Now suppose that κ *is very small*, though still just greater than zero; to be more precise suppose that the ratio κ/β is negligible, so that while immigration is sufficient to start off the process, and to restart it if ever the population is wiped out by an excess of the μ-effect, it is negligible when compared with the contributions from the β-effect while the colony is actually growing. Then, exactly as in Fisher's analysis referred to in § 1, it will be seen that the size distribution of such a colony observed at any time will be a logarithmic series: in fact the distribution of n, *given that* $n > 0$, is*

$$\frac{x^n}{ny} \quad (n = 1, 2, 3, \ldots),$$

where as always $y = -\ln(1-x)$, and

$$x = \frac{\beta(\Lambda - 1)}{\beta\Lambda - \mu}. \tag{21}$$

If T is very large, then $$x \simeq 1 - \left(1 - \frac{\mu}{\beta}\right)\frac{1}{\Lambda} \quad \text{when} \quad \beta > \mu, \tag{22}$$

and $$x \simeq \beta/\mu \quad \text{when} \quad \beta < \mu. \tag{23}$$

The second case is the more interesting, for one can then let T tend to infinity and so obtain the stable distribution of population size when $\beta < \mu$. Of course in the limit when $\kappa = 0$ it is 'almost certain' that $n = 0$. If, however, a colony *does* exist (i.e. if $n > 0$), then it is almost certainly homogeneous (descended from a single immigrant ancestor), and its size n will be distributed in a logarithmic series.

(iii) The case $\beta = \mu$, when reproduction and mortality just balance, requires a separate discussion. The equations auxiliary to (12) are now

$$-dt = \frac{dz}{\beta(z-1)^2} = -\frac{d\phi}{\kappa(z-1)\phi},$$

and the general integral is $$\phi(z, t) = (1-z)^{-\kappa/\beta}\Phi\left\{\beta t + \frac{1}{1-z}\right\}.$$

Condition (A) then gives

$$\phi(z, 0) = (1+\beta T)^{-\kappa/\beta}\left\{1 - \frac{\beta T}{1+\beta T}z\right\}^{-\kappa/\beta}, \tag{24}$$

so that here again n is distributed as a negative binomial variate with index κ/β, but the mean value is now

$$E(n) = \kappa T, \tag{25}$$

so that the expected size of the population is linearly proportional to the time of exposure to immigration. If κ/β is very small the limiting conditional distribution of n (given that $n > 0$) is once again the Fisher series

$$\frac{x^n}{ny} \quad (n = 1, 2, 3, \ldots),$$

where now $$x = \frac{\beta T}{1+\beta T}. \tag{26}$$

* This conditional distribution is of course obtained by taking the ratio of the general term to the sum of all the terms but the first, in the negative binomial series, and then letting $\kappa/\beta \to 0$.

5. It will perhaps have been noticed that in the analysis of the last section the immigration rate κ (when κ/β is small) merely has the effect of ensuring that an observed population is almost certainly descended from a single immigrant ancestor who entered the region considered at some instant during the time interval of length T preceding the moment of observation. I think this gives us a clue to the true 'explanation' of the occurrence of the logarithmic series in the solution to the problem just considered. Before exploring the matter further it will be found helpful to examine the growth of a single family by setting $\kappa = 0$ and starting from unit population at time $t = -T$, thus employing boundary condition (B).

From the generating function (18) it then follows that the distribution of the population size n at the time of observation ($t = 0$) is*

$$P_0 = \frac{\mu(\Lambda-1)}{\beta\Lambda-\mu} \quad \text{and} \quad P_n = (1-P_0)(1-u)u^{n-1} \quad (n>0), \tag{27}$$

where

$$u = \frac{\beta(\Lambda-1)}{\beta\Lambda-\mu}.$$

The distribution is thus a geometric series with a modified zero term, the mean population size being

$$E(n) = \Lambda,$$

so that the *expected* population grows geometrically at an exponential rate $\beta-\mu$ (which will, of course, be negative if $\beta<\mu$).

It is of interest to evaluate the variance of the population size. This proves to be

$$\text{Var}(n) = E(n-\Lambda)^2 = \frac{\beta+\mu}{\beta-\mu}\Lambda(\Lambda-1), \tag{28}$$

and the coefficient of variation of the population size is thus

$$\sqrt{\left(\frac{\beta+\mu}{\beta-\mu}\right)}\sqrt{\left(1-\frac{1}{\Lambda}\right)} \quad \text{if} \quad \beta>\mu$$

and

$$\sqrt{\left(\frac{\mu+\beta}{\mu-\beta}\right)}\sqrt{\left(\frac{1}{\Lambda}-1\right)} \quad \text{if} \quad \beta<\mu.$$

For large T these expressions become

$$\text{C. of V.}(n) \simeq \sqrt{\left(\frac{\beta+\mu}{\beta-\mu}\right)} \quad \text{if} \quad \beta>\mu, \tag{29}$$

and

$$\text{C. of V.}(n) \simeq \sqrt{\left\{\frac{\mu+\beta}{\mu-\beta}\frac{1}{E(n)}\right\}} \quad \text{if} \quad \beta<\mu; \tag{30}$$

in the second case it will be recalled that $E(n)\to 0$ as $T\to\infty$.

The distribution of the population size thus behaves rather differently in the two cases of an exponentially growing and an exponentially decreasing population; this agrees with a conclusion reached by M. S. Bartlett (1937). He considered the similar problem when there is no spreading of generations and the generation-time is rigorously constant. As he points out, there is a connexion with Fisher's theory of the extinction of rare characters and, one might add, with the work of Francis Galton, H. W. Watson, and A. J. Lotka on the extinction of surnames.†

* (*Added in proof.*) Dr Bartlett has pointed out to me that the result (27) is stated by N. Arley & V. Borchsenius in *Acta Math.* (1945), **76**, 298–9. It is attributed by them to Dr C. Palm.

† See Fisher (1930), Galton (1889) and Lotka (1931). The problem of the distribution of surnames, and its variation in time, seems not yet to have received all the attention it deserves. The surname is a 'rare character' whose extinction can very readily be observed; normal social conditions ensure that it is inherited as if it were controlled by a gene totally sex-linked in Y. Reference may be made to the work of R. A. Fisher & Janet Vaughan (1939), and J. A. Fraser Roberts (1941–2), who have considered the relation between surnames and blood-groups. [See also reference to Tocher & Gray, p. 1 above. Ed.]

When $\beta = \mu$, so that the forces of reproduction and mortality just balance, the above solution must be modified. The appropriate results are most easily obtained by letting β tend to μ in the several formulae. In this way one finds

$$P_0 = \frac{\beta T}{1+\beta T} \quad \text{and} \quad P_n = (1-P_0)(1-u)u^{n-1} \quad (n>0), \tag{31}$$

where

$$u = \frac{\beta T}{1+\beta T},$$

and so

$$E(n) = 1, \quad \text{Var}(n) = 2\beta T \quad \text{and} \quad \text{C. of V.}(n) = \sqrt{(2\beta T)}. \tag{32}$$

It is of interest to note that if β and μ instead of being constants are each proportional to the *same* function of the time,* the above theory still holds, provided that T is everywhere replaced by

$$\int_{-T}^{0} \psi(t)\,dt,$$

where $\beta = \beta_0\psi(t)$ and $\mu = \mu_0\psi(t)$.

6. Consider now the size distribution of a colony developing in the absence of immigration and known to have originated from a single individual whose arrival in the region concerned occurred during the preceding T time-units. If the time of arrival of the common ancestor is a random variable uniformly distributed from $t = -T$ to $t = 0$, it will follow from (27) that

$$P_n = \frac{1}{\beta T}\int_0^T \left\{\frac{\beta(\lambda-1)}{\beta\lambda-\mu}\right\}^{n-1} \frac{\beta\lambda(\beta-\mu)^2}{(\beta\lambda-\mu)^2}\,d\tau \quad (n \geqslant 1),$$

where $\lambda = e^{(\beta-\mu)\tau}$. Now since

$$\frac{d}{d\tau}\left\{\frac{\beta(\lambda-1)}{\beta\lambda-\mu}\right\} = \frac{\beta\lambda(\beta-\mu)^2}{(\beta\lambda-\mu)^2},$$

this can be written

$$P_n = \frac{1}{\beta T}\int_0^U u^{n-1}\,du = \frac{U^n}{n\beta T},$$

while

$$P_0 = 1 + \frac{1}{\beta T}\ln(1-U),$$

U being the value of

$$u = \frac{\beta(\lambda-1)}{\beta\lambda-\mu}$$

when $\tau = T$. Thus the distribution of the population size at the time of observation ($t = 0$) is

$$P_0 = 1 - \frac{1}{\beta T}\ln\left(\frac{1}{1-x}\right) \quad \text{and} \quad P_n = (1-P_0)\frac{x^n}{ny} \quad (n>0), \tag{33}$$

where

$$x = \frac{\beta(\Lambda-1)}{\beta\Lambda-\mu}, \tag{34}$$

y and Λ being defined as before. For large T,

$$x \simeq 1 - \left(1-\frac{\mu}{\beta}\right)\frac{1}{\Lambda} \quad \text{if} \quad \beta > \mu, \tag{35}$$

and

$$x \simeq \beta/\mu \quad \text{if} \quad \beta < \mu. \tag{36}$$

In the latter case it is permissible to let T tend to infinity and so obtain the stable (logarithmic series) distribution for n (*given that* $n>0$). When $\beta = \mu$ the (x, T) relation is $x = \beta T/(1+\beta T)$.

It will now be clear why, in § 4 (ii), the boundary condition (A) together with the hypothesis $n>0$ led to a logarithmic series distribution for n as $\kappa/\beta \to 0$. For in these circumstances it

* A discussion of the similar problem when β and μ are *any* (not necessarily the same) functions of the time will be given in my paper (1948).

would be almost certain that an observed population was wholly descended from a single immigrant ancestor who arrived in the field of observation at an unknown time instant uniformly distributed between $t = -T$ and $t = 0$.*

7. The classical theory of population growth (see, for example, A. J. Lotka (1945) for a general review and extensive references) is largely based on a *deterministic* description of the phenomena, which leads to differential and integral equations for the expectation values of the random variables concerned (the total population size, and the numbers in the several age groups). Apart from the work on 'extinction' already mentioned, the first *stochastic* treatment of the general problems of population growth seems to be that of W. Feller in an important paper† which is unfortunately not generally accessible in this country. Feller's work has been further developed by N. Arley (1943),‡ who showed that discontinuous Markoff processes of a similar type are equally relevant in the theory of the 'cascade showers' initiated by cosmic ray particles. In particular he makes use of the simple birth-and-death process discussed here in § 5, and quotes Feller's formulae for the mean and variance of n as functions of the time t. Arley gives an elegant method for determining an expansion for $P_n(t)$ in powers of $(t-t_0)$, and observes that the calculation of high order coefficients becomes very cumbersome. One can now see, on examining the complete solution contained in (27), that this is very reasonable, for $P_n(t)$ is quite a complicated function of the time t, although a simple one of the reduced variable u.

The results of the present paper would not have been obtained without the generating-function technique for transforming the differential-difference equations into a partial differential equation of simple type. This device was suggested to me by Dr M. S. Bartlett in the summer of 1946, and he has also applied his method to a number of Markoff processes of interest in biology, one of which is the birth-and-death process of § 5. An account of this work is now available (Bartlett, 1947). In another place I intend to give a discussion of the most general birth-and-death process of this type, in which the birth- and death-rates β and μ can be any functions of the time t (Kendall, 1948); this development also makes use of the generating-function procedure.

It is a pleasure to acknowledge my indebtedness to Dr Bartlett and to the many other friends whose comments have helped to clarify my ideas on this subject. In particular, I should like to thank Mr D. J. Finney and members of his Seminar in Biological Statistics in Oxford, and Mr F. J. Anscombe, Dr P. Jones and Dr C. B. Williams of the Rothamsted Experimental Station.

* The condition $n > 0$ is to be introduced *after* (33) has been established. If it were imposed from the start, the *a posteriori* distribution of the time of arrival of the immigrant ancestor would no longer be a uniform one, but the appropriate modification of the argument would lead to the same final result.

† Feller (1939). I am greatly indebted to Prof. Feller for the gift of a reprint of this paper. In private correspondence he tells me that he also, in unpublished lectures, has solved the equations governing the birth-and-death process discussed here in § 5.

‡ Arley's monograph is also of great value in presenting a useful account (including several new developments) of the Kolmogoroff-Feller theory of Markoff processes, especially those of the discontinuous type used here. Another excellent account has been given recently by O. Lundberg (1940). For a general introduction to this subject, reference may be made to my review article (1947).

REFERENCES

ARLEY, N. (1943). *On the Theory of Stochastic Processes and their Application to the Theory of Cosmic Radiation*. Copenhagen: G. E. C. Gads.

BARTLETT, M. S. (1937). Deviations from expected frequencies in the theory of inbreeding. *J. Genet.* **35**, 83–7.

BARTLETT, M. S. (1947). *Stochastic Processes*. (Notes of a course given at the University of North Carolina in the Fall Quarter, 1946. It is understood that copies of these notes are available on request.)

FELLER, W. (1939). Die Grundlagen der Volterraschen Theorie des Kampfes ums Dasein in Wahrscheinlichkeitstheoretischer Behandlung. *Acta Biotheoretica*, **5**, 11–40. This seems not to be generally available in Great Britain, but there is an abstract in *Math. Rev.* **1**, 22.

FISHER, R. A. (1930). *The Genetical Theory of Natural Selection*, pp. 73–83. Oxford: Clarendon Press.

FISHER, R. A. & VAUGHAN, JANET (1939). Surnames and blood-groups. *Nature, Lond.*, **144**, 1047.

FISHER, R. A., CORBET, A. S. & WILLIAMS, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12**, 42–58.

GALTON, FRANCIS (1889). *Natural Inheritance*, Appendix F (which includes the contribution by H. W. WATSON). London: Macmillan.

GREENWOOD, M. & YULE, G. U. (1920). *J. R. Statist. Soc.* **83**, 255–79.

HINSHELWOOD, C. N. (1946). *The Chemical Kinetics of the Bacterial Cell*, Chapter x (see especially Fig. 64). Oxford: University Press.

KELLY, C. D. & RAHN, O. (1932). The growth rate of individual bacterial cells. *J. Bacteriol.* **23**, 147–53.

KENDALL, D. G. (1947). A review of some recent work on discontinuous Markoff processes with applications to biology, physics, and actuarial science. *J.R. Statist. Soc.* **110**, 130–7.

KENDALL, D. G. (1948). On the generalized 'birth-and-death' process. *Ann. Math. Statist.* (in the Press).

LOTKA, A. J. (1931). The extinction of families. *J. Wash. Acad. Sci.* **21**, 377–80 and 453–9.

LOTKA, A. J. (1945). Population analysis as a chapter in the mathematical theory of evolution. Included in *Essays on Growth and Form presented to D'Arcy Wentworth Thompson* (edited by Le Gros Clark, W. E. & Medawar, P. B.). Oxford: Clarendon Press.

LUNDBERG, O. (1940). *On Random Processes and their Application to Sickness and Accident Statistics*. Uppsala: Almqvist and Wiksells.

ROBERTS, J. A. FRASER (1941–2). Blood-group frequencies in north Wales. *Ann. Eugen., Lond.*, **11**, 260–71. (Includes references to earlier work on the distribution of surnames.)

WILLIAMS, C. B. (1944*a*). Some applications of the logarithmic series and the index of diversity to ecological problems. *J. Ecol.* **32**, 1–44.

WILLIAMS, C. B. (1944*b*). The numbers of publications written by biologists. *Ann. Eugen., Lond.*, **12**, 143–6.

WILLIAMS, C. B. (1947). The logarithmic series and its application to biological problems. *J. Ecol.* (in the Press).

THE STUDENTIZED FORM OF THE EXTREME MEAN SQUARE TEST IN THE ANALYSIS OF VARIANCE†

By K. R. NAIR

CONTENTS

	PAGE
Part I. Expansion of the studentized integral: extension of H. O. Hartley's results	
1. Introduction	16
2. Expansion of the studentized integral in powers of ν^{-1}	17
3. Application to the incomplete beta-function	19
4. Application to 'Student's' integral	20
5. Some further applications	20
Part II. Application to tests regarding the largest and smallest of several variances	
1. Introduction	21
2. Probability integral of the largest variance ratio	22
3. Construction of tables	24
4. An illustration	26
5. Probability integral of the smallest variance ratio	28

Part I. Expansion of the studentized integral: extension of H. O. Hartley's results

1. *Introduction*

Let $x_1, \ldots, x_n$ denote a sample of n observations drawn from a normal population with mean μ and standard deviation σ. Then the difference between sample and population mean, $\bar{x}-\mu$, is normally distributed with zero mean and standard deviation $\sigma/\sqrt{n}$. If we estimate the standard deviation of the parent, σ, by

$$s = \sqrt{[\Sigma(x-\bar{x})^2/(n-1)]},$$

then 'Student' (1908) gave the distribution of

$$t = (\bar{x}-\mu)\sqrt{n}/s.$$

As was to be expected, this distribution was independent of σ. The knowledge of the distribution of t made it possible to draw inferences, with the help of evidence entirely supplied by the sample, about the location-parameter μ of a normal population, without making any assumptions about the generally unknown scale-parameter, σ.

Neyman & Pearson (1928) extended this notion to an analogous problem connected with the rectangular and exponential populations.

Fisher (1924) obtained the distribution of s'/s where s' and s are two independent estimates of σ, calculated by the root-mean-square method. This distribution is also independent of σ. Fisher also showed that 'Student's' t-distribution could be extended to a wide range of sampling problems. Sukhatme (1937) has shown that analogues to t and s'/s of 'normal theory' can be developed for an exponential population, again eliminating the unknown scale parameter.

This notion of eliminating unknown scale parameters from the distribution laws of statistics has come to be known as 'studentization'.

† Part of a thesis approved for the degree of Ph.D. of the University of London.

If instead of s' and s which give unbiased estimates of σ, we use other types of statistics which are 'proportional' to σ,* such as the range, mean deviation, etc., it is clear that the distribution of the ratio of two such estimates will be independent of σ, although the analytic expression for this distribution may be difficult to obtain.

Hartley (1938, 1944) has shown that if instead of s'/s we have a ratio w/s, where (i) w is a general statistic 'proportional' to σ calculated from a sample of n observations $x_1, \ldots, x_n$ drawn from a normal population and (ii) s is independently distributed with ν degrees of freedom, then the distribution of w/s can be derived without much difficulty if the distribution of w/σ is known. His solution is described in some detail in the next section and its application to special problems is considered in the succeeding sections.

The probability integral of w/s may be called the *studentized integral* of w. When $\nu \to \infty$ the studentized integral becomes identical with the probability integral of w/σ, which may in turn be called the ∞-*integral* of w/s.

2. *Expansion of the studentized integral in powers of* ν^{-1}

Let $f_\nu(w/s)$ denote the distribution function of w/s. When $\nu \to \infty$ this gives the distribution function of w/σ, which we shall denote by $f(w/\sigma)$.

Let the probability integral of w/s be

$$P_\nu(Q) = \int_0^Q f_\nu(w/s)\, d(w/s).$$

When $\nu \to \infty$, $P_\nu(Q)$ gives the probability integral of w/σ which we may denote as $P(Q)$.

Assuming σ to be unity, it follows (see, for example, Hartley) that

$$P_\nu(Q) = 2\Gamma(\tfrac{1}{2}\nu)^{-1}(\tfrac{1}{2}\nu)^{\frac{1}{2}\nu}\int_0^\infty s^{\nu-1}e^{-\frac{1}{2}\nu s^2}P(Qs)\,ds. \tag{1}$$

Hartley made the significant discovery that a recurrence formula connecting $P_\nu(Q)$ and $P_{\nu-2}(Q)$ could be built up and that, thereby, $P_\nu(Q)$ can be obtained as the solution of a partial differential equation. He solves this equation by iteration, and at the third stage obtains the first three terms in an expansion for $P_\nu(Q)$ in powers of ν^{-1}, namely,

$$P_\nu(Q) = P(Q) + \frac{1}{4\nu}(Q^2P'' - QP') + \frac{1}{6(4\nu)^2}(3Q^4P^{\mathrm{iv}} - 2Q^3P''' - 3Q^2P'' + 3QP'), \tag{2}$$

where the derivatives of P are taken at the argument Q.

For ν of the order 20 or larger this formula will in most cases give quite sufficient accuracy. Sometimes when ν is small, however, we may need at least the next two terms of the expansion. The term in ν^{-3} immediately follows from equation (24) of Hartley's (1944) paper and is

$$\frac{e^{-3x}}{12}(\tfrac{1}{32}\phi^{\mathrm{vi}} - \tfrac{7}{16}\phi^{\mathrm{v}} + \tfrac{17}{8}\phi^{\mathrm{iv}} - \tfrac{17}{4}\phi''' + 3\phi''), \tag{3}$$†

where $$\phi(\lambda) = P(e^\lambda) = P(Q) \quad \text{and} \quad x = \log \nu.$$

The term in ν^{-4} will come from the fifth stage of the iteration which involves some lengthy algebra. In terms of x, ϕ and λ this becomes

$$\frac{e^{-4x}}{360(4^4)}(15\phi^{\mathrm{viii}} - 360\phi^{\mathrm{vii}} + 3320\phi^{\mathrm{vi}} - 14784\phi^{\mathrm{v}} + 32240\phi^{\mathrm{iv}} - 28800\phi''' + 1280\phi'' + 6144\phi'). \tag{4}$$

* A statistic w is 'proportional' to σ if it transforms into w/σ when the observations x_i are measured in units of σ.

† The coefficient of ϕ''' is misprinted $-\frac{11}{6}$ for $-\frac{17}{4}$ in Hartley's paper.

The derivatives of $\phi(\lambda)$ can be expressed in terms of those of $P(Q)$.

$$\left.\begin{aligned}
\phi(\lambda) &= P(Q),\\
\phi' &= QP',\\
\phi'' &= Q^2P'' + QP',\\
\phi''' &= Q^3P''' + 3Q^2P'' + QP',\\
\phi^{\mathrm{iv}} &= Q^4P^{\mathrm{iv}} + 6Q^3P''' + 7Q^2P'' + QP',\\
\phi^{\mathrm{v}} &= Q^5P^{\mathrm{v}} + 10Q^4P^{\mathrm{iv}} + 25Q^3P''' + 15Q^2P'' + QP',\\
\phi^{\mathrm{vi}} &= Q^6P^{\mathrm{vi}} + 15Q^5P^{\mathrm{v}} + 65Q^4P^{\mathrm{iv}} + 90Q^3P''' + 31Q^2P'' + QP',\\
\phi^{\mathrm{vii}} &= Q^7P^{\mathrm{vii}} + 21Q^6P^{\mathrm{vi}} + 140Q^5P^{\mathrm{v}} + 350Q^4P^{\mathrm{iv}} + 301Q^3P''' + 63Q^2P'' + QP',\\
\phi^{\mathrm{viii}} &= Q^8P^{\mathrm{viii}} + 28Q^7P^{\mathrm{vii}} + 266Q^6P^{\mathrm{vi}} + 1050Q^5P^{\mathrm{v}} + 1701Q^4P^{\mathrm{iv}} + 966Q^3P'''\\
&\quad + 127Q^2P'' + QP'.
\end{aligned}\right\} \tag{5}$$

Substituting these in (3) and (4) we get, up to terms in ν^{-4},

$$P_\nu(Q) = a_0 + a_1/\nu + a_2/\nu^2 + a_3/\nu^3 + a_4/\nu^4, \tag{6}$$

where

$$\left.\begin{aligned}
a_0 &= P(Q),\\
a_1 &= \tfrac{1}{4}(Q^2P'' - QP'),\\
a_2 &= \frac{1}{6(4^2)}(3Q^4P^{\mathrm{iv}} - 2Q^3P''' - 3Q^2P'' + 3QP'),\\
a_3 &= \frac{1}{6(4^3)}(Q^6P^{\mathrm{vi}} + Q^5P^{\mathrm{v}} - 7Q^4P^{\mathrm{iv}} + 12Q^3P''' - 15Q^2P'' + 15QP'),\\
a_4 &= \frac{1}{360(4^4)}(15Q^8P^{\mathrm{viii}} + 60Q^7P^{\mathrm{vii}} - 250Q^6P^{\mathrm{vi}} + 366Q^5P^{\mathrm{v}},\\
&\qquad - 285Q^4P^{\mathrm{iv}} - 30Q^3P''' + 945Q^2P'' - 945QP').
\end{aligned}\right\} \tag{7}$$

From (7) we can derive the following results

$$\left.\begin{aligned}
&\int QP'\,dQ = -\frac{4}{3}\int a_1\,dQ = -\frac{32}{25}\int a_2\,dQ = -\frac{128}{105}\int a_3\,dQ = -\frac{2048}{1659}\int a_4\,dQ,\\
&\int Q^2P'\,dQ = -\int Qa_1\,dQ = -\frac{1}{2}\int Qa_2\,dQ = -\frac{1}{4}\int Qa_3\,dQ = -\frac{1}{8}\int Qa_4\,dQ,\\
&\int Q^3P'\,dQ = -\frac{4}{5}\int Q^2a_1\,dQ = -\frac{96}{385}\int Q^2a_2\,dQ = -\frac{128}{1575}\int Q^2a_3\,dQ = -\frac{2048}{76153}\int Q^2a_4\,dQ,\\
&\int Q^4P'\,dQ = -\frac{2}{3}\int Q^3a_1\,dQ = -\frac{1}{7}\int Q^3a_2\,dQ = -\frac{1}{30}\int Q^3a_3\,dQ = -\frac{1}{124}\int Q^3a_4\,dQ,
\end{aligned}\right\} \tag{8}$$

where the limits of integration are 0 and ∞.

These results enable us to express the moments of the studentized statistic in terms of those of the non-studentized statistic. Thus, up to terms in ν^{-4}, we have

$$\left.\begin{aligned}
\int_0^\infty P_\nu(Q)\,dQ &= \int_0^\infty P(Q)\,dQ - \left(\frac{3}{4\nu} + \frac{25}{32\nu^2} + \frac{105}{128\nu^3} + \frac{1659}{2048\nu^4}\right)\int_0^\infty QP'(Q)\,dQ,\\
\int_0^\infty QP_\nu(Q)\,dQ &= \int_0^\infty QP(Q)\,dQ - \left(\frac{1}{\nu} + \frac{2}{\nu^2} + \frac{4}{\nu^3} + \frac{8}{\nu^4}\right)\int_0^\infty Q^2P'(Q)\,dQ,\\
\int_0^\infty Q^2P_\nu(Q)\,dQ &= \int_0^\infty Q^2P(Q)\,dQ - \left(\frac{5}{4\nu} + \frac{385}{96\nu^2} + \frac{1575}{128\nu^3} + \frac{76153}{2048\nu^4}\right)\int_0^\infty Q^3P'(Q)\,dQ,\\
\int_0^\infty Q^3P_\nu(Q)\,dQ &= \int_0^\infty Q^3P(Q)\,dQ - \left(\frac{3}{2\nu} + \frac{7}{\nu^2} + \frac{30}{\nu^3} + \frac{124}{\nu^4}\right)\int_0^\infty Q^4P'(Q)\,dQ.
\end{aligned}\right\} \tag{9}$$

3. *Application to the incomplete beta-function*

By applying expansion (2) to the probability integral of the square root of a variance ratio having $2p$ and $2q$ degrees of freedom, Hartley obtained a new formula for the incomplete beta-function

$$I_x(p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\,\Gamma(q)}\int_0^x x^{p-1}(1-x)^{q-1}\,dx$$

in terms of the incomplete gamma-function and its derivatives.

Using the additional terms in ν^{-3} and ν^{-4} given in (6) this formula becomes

$$I_x(p,q) = \Gamma(p)^{-1}\int_0^\omega e^{-y}y^{p-1}dy + \Gamma(p)^{-1}e^{-\omega}\omega^p\{b_1/(2q)+b_2/(2q)^2+b_3/(2q)^3+b_4/(2q)^4\}, \quad (10)$$

where $x = \omega/(\omega+q)$ and

$$\left.\begin{aligned}
b_1 &= \{(p-1)-\omega\},\\
b_2 &= \tfrac{1}{6}\{(p-1)(p-2)(3p-1)-(p-1)(9p-2)\omega+(9p-1)\omega^2-3\omega^3\},\\
b_3 &= \tfrac{1}{6}\{p(p-1)^2(p-2)(p-3)-p(p-1)(p-2)(5p-3)\omega+2p(p-1)(5p-1)\omega^2\\
&\quad -2p(5p+1)\omega^3+(5p+3)\omega^4-\omega^5\},\\
b_4 &= \tfrac{1}{360}\{(p-1)(p-2)(p-3)(p-4)(15p^3-30p^2+5p+2)\\
&\quad -(p-1)(p-2)(p-3)(105p^3-135p^2+10p+8)\omega\\
&\quad +(p-1)(p-2)(315p^3-180p^2+5p+12)\omega^2\\
&\quad -(p-1)(525p^3+75p^2+50p+8)\omega^3\\
&\quad +(525p^3+450p^2+175p+2)\omega^4\\
&\quad -5(63p^2+99p+46)\omega^5+15(7p+9)\omega^6-15\omega^7\}.
\end{aligned}\right\} \quad (11)$$

This formula strictly applies only when p and q of the incomplete beta-function are such that $2p$ and $2q$ are positive integers. It is not suitable when p is large, because of the slow convergence of $b_i/(2q)^i$. For small values of p and moderate or large values of q, however, the formula is quite useful. For large p and q, Wishart (1927) and others have developed useful methods.

To give some idea of the accuracy of the formula when only terms up to b_2 are used, Hartley considered the example $p = 1$, $\omega = \frac{5}{9}$ and various integral values of q in the range 5 to 50. The agreement between the exact and the approximate values was very good for values of $q > 20$. For lower values of q, addition of the terms in b_3 and b_4 makes the approximations closely agree with the exact values.

Table giving exact and approximate values of $I_x(p,q)$ for $p = 1$, $\omega = \frac{5}{9}$ and $x = \omega/(\omega+q)$

q	Exact	Approximation		
		(1)	(2)	(3)
5	0·976 155	0·974 627	0·975 597	0·976 788
10	0·987 945	0·987 774	0·987 896	0·987 970
15	0·991 140	0·991 093	0·991 129	0·991 144
20	0·992 571	0·992 553	0·992 568	0·992 572

In this table approximation (1) uses terms up to b_2, and approximations (2) and (3) are obtained by including terms up to b_3 and b_4 respectively. The last approximation gives values in *excess* of the exact values, which shows that the value of the omitted remainder terms for $\omega = \frac{50}{9}$ is negative, though negligible when q exceeds 10.

4. *Application to 'Student's' integral*

Fisher (1926) gave an expansion for the probability integral of 'Student's' t, in powers of ν^{-1}. He worked out the coefficients of the terms up to ν^{-5} and found that the maximum value of the fifth correction never exceeded 10^{-5} when ν was greater than 18. It will be interesting to compare the coefficients in Fisher's expansion with those obtained by applying Hartley's method.

Since t is symmetrically distributed about 0, we may consider the probability integral of $|t|$. To get the expansion of this integral by Hartley's method we have only to replace the $P(Q)$ in (6) by the '∞-integral' of $|t|$, namely,

$$P(Q) = \sqrt{\frac{2}{\pi}} \int_0^Q e^{-\frac{1}{2}t^2} d|t|. \tag{12}$$

Alternatively, since $|t|$ is the same as the square root of a variance ratio having 1 and ν degrees of freedom, putting $p = \frac{1}{2}$, $q = \frac{1}{2}\nu$ and $\omega = \frac{1}{2}t^2$ in (10), we may obtain its probability integral in the form

$$\sqrt{\frac{2}{\pi}} \int_0^{|t|} e^{-\frac{1}{2}x^2} dx - \sqrt{\frac{2}{\pi}} |t| e^{-\frac{1}{2}t^2} \sum_{r=1}^{4} T_r \nu^{-r}, \tag{13}$$

where
$$\left.\begin{aligned} T_1 &= \tfrac{1}{4}(t^2+1), \\ T_2 &= \frac{1}{6(4^2)}(3t^6 - 7t^4 - 5t^2 - 3), \\ T_3 &= \frac{1}{6(4^3)}(t^{10} - 11t^8 + 14t^6 + 6t^4 - 3t^2 - 15), \\ T_4 &= \frac{1}{360(4^4)}(15t^{14} - 375t^{12} + 2225t^{10} - 2141t^8 - 939t^6 - 213t^4 + 915t^2 + 945). \end{aligned}\right\} \tag{14}$$

The probability integral of t from t to ∞ immediately follows as

$$\frac{1}{\sqrt{(2\pi)}} \int_t^\infty e^{-\frac{1}{2}x^2} dx + \frac{1}{\sqrt{(2\pi)}} t e^{-\frac{1}{2}t^2} \sum_{r=1}^{4} T_r \nu^{-r}. \tag{15}$$

The coefficients T_1, T_2, T_3 and T_4* are identical to those given by Fisher providing us with a useful confirmation of the general applicability of Hartley's method.

5. *Some further applications*

In his 1938 paper, Hartley examined the studentized probability integral of the largest and smallest among k independent estimates of a variance, each having 1 degree of freedom. The importance of this problem in the analysis of factorial experiments was first pointed out by Wishart (1938). As an expansion of the integral in powers of ν^{-1} was not then known, Hartley approached the problem by an approximate method which was satisfactory for getting the lower percentage points of the smallest variance, but not for the more important case of upper percentage points of the largest variance.

* The coefficient of t^2 in T_4 appears with a negative sign in Fisher's (1926) paper but the correct sign has been used in a later paper, viz. *Ann. Eugen.* **11**, 141–72.

Meanwhile, Finney (1941) gave an exact solution for the studentized integral of the largest and smallest of k variances, each estimated with m degrees of freedom. It is not easy to make this exact solution amenable to numerical calculation of the percentage points for odd values of m. Hartley's expansion (2) with the additional terms given in (6) could usefully be employed to get the percentage points when $m = 1$. Details of the method and tables of the 5 and 1 % points are given in Part II.

Pearson & Hartley (1943) used expansion (2) to calculate the studentized integral of the range. This is the probability integral of $(x_n - x_1)/s$, where x_1 and x_n are the two extreme observations in a sample of n observations and s is an independent estimate of σ based on ν degrees of freedom. The studentized range is of importance in quality control charts for industrial products. A closely allied statistic, namely, the studentized extreme deviate $(x_n - \bar{x})/s$ (or $(\bar{x} - x_1)/s$) has useful applications in designed experiments when we are concerned with judging whether a single outlying treatment (best or worst) really differs from the rest. The probabilit integral of this statistic will be considered in another paper (Nair, 1948) and tables provided for $n = 3, 4, \ldots, 9$.

There are many other examples where the expansion of the studentized integral can usefully be employed.The applications cited above are convincing proof of its potentialities.

PART II. APPLICATION TO TESTS REGARDING THE LARGEST AND SMALLEST OF SEVERAL VARIANCES

1. *Introduction*

Let $s_1^2, \ldots, s_k^2$ be independent estimates of an unknown variance σ^2 of a normal population each calculated from sums of squares having m degrees of freedom and arranged in ascending order of magnitude. Let s_0^2 be another independent estimate based on ν degrees of freedom against which each of the first k estimates is to be tested for significant differences from s_0^2. In routine analysis of variance where such tests frequently occur, it is customary to test each of the k variances against s_0^2 and separately declare whether there is a significant difference or not. Forming now for each of the independent s_i^2 the ratio $F_i = s_i^2/s_0^2$, it is obvious that the largest ratio, $F_k = s_k^2/s_0^2$, is more likely to be declared significant than any other variance ratio in the set. Indeed, for $k = 20$ it would be expected to be significant at the 5 % level of F for m and ν degrees of freedom. This source of bias can be eliminated if the probability integral of the largest variance ratio is numerically evaluated.

Although the theory discussed in this paper applies to any value of m, owing to practical difficulties, tables of the upper 5 and 1 % points of the largest variance ratio have been prepared only when $m = 1$, which is the most important case in practice. Thus, as Wishart (1938) has pointed out, it has useful application in the analysis of variance for $2 \times 2 \times 2 \times \ldots$ factorial design. It could in fact be applied to the analysis of any designed experiment where the total variance due to treatment effects is split up into components each having a single degree of freedom.

Another possible application is in the fitting of curves with orthogonal terms, e.g. orthogonal polynomials and harmonic analysis. It may well happen that while, taken as a whole there is no significant curvilinearity, the coefficient of a single high order term turns out to be very large compared to the rest. The new tables will be useful in testing whether such isolated coefficients are significantly large.

The theory developed for the largest variance ratio could easily be extended to any other *ranked* ratio. Of these the smallest, namely, $F_1 = s_1^2/s_0^2$ has been considered. Occasions to test the significance of the smallest of k variance ratios are much less frequent than that of the largest. In the former case, we are concerned with the *lower* percentage points of the probability integral, in order that a test could be made whether an observed smallest ratio is significantly small.

In field experiments, the smallest of k variances tested may become significantly small compared to the 'error' variance (s_0^2) if the $m+1$ groups of plot values from which the m degrees of freedom of that variance were obtained showed a high negative intra-class correlation. The extreme value that this correlation can have, if there are l plots in each group, is $-1/(l-1)$, so that it can be -1 only if $l = 2$. Wishart (1938) again gives an example where a test for significance of smallest variance ratio when $m = 1$ is appropriate.

A rather interesting use for the test of significance of the smallest variance ratio, when $m = 1$, can be found in certain methods of statistical control in sample surveys devised by Mahalanobis (1944). If, say, you are sampling agricultural fields in k districts and send two investigators to each district to collect half the quota of sample units into two randomly selected half samples A and B, a comparison of the A and B mean values for each district should give a clue as to whether the data have been properly collected. If one of the A-B differences is very large and the variance ratio significant it is usual to conclude that the two investigators concerned did not carry out the instructions in the same manner or that some other personal bias had crept in. On the other hand if one of the k differences between (A, B) pairs is surprisingly small and turns out to give a significant smallest variance ratio, it may perhaps arouse a suspicion that the two investigators consulted each other and dishonestly made the means of their data agree. This is the negative side of the argument in favour of a test of significance of the smallest variance ratio. On the positive side, it may often turn out that the smallest variance ratio is *not* significantly small, avoiding awkward aspersions on the reliability of the investigators.

2. *Probability integral of the largest variance ratio*

Let ${}_\nu P_k(Q)$ denote the probability of s_k/s_0 being $\leqslant Q$. This is the same as the probability of $F_k = s_k^2/s_0^2$ being $\leqslant Q^2$. When $\nu \to \infty$, this probability will be denoted by $P_k(Q)$.

Finney (1941) showed that

$$ {}_\nu P_k(Q) = M^k(1+\lambda)^{-\frac{1}{2}\nu}, \tag{1} $$

where

$$ M = \Gamma(\tfrac{1}{2}m)^{-1}\int_0^{-\frac{m}{\nu}Q^2\frac{\partial}{\partial\lambda}} u^{\frac{1}{2}m-1}e^{-u}\,du, \tag{2}* $$

and λ is put equal to zero in (1) after differentiation.

He found that (1) was not amenable to numerical calculations if m is odd. The simplest case is $m = 2$ for which the probability integral reduces to

$$ {}_\nu P_k(Q) = \sum_{r=0}^{k}(-1)^r\,{}^kC_r\left(1+\frac{2rQ^2}{\nu}\right)^{-\frac{1}{2}\nu}. \tag{3} $$

Hartley (1938) had suggested that if we assume the k variance ratios to be independent, an assumption quite justifiable if ν is very large, ${}_\nu P_k(Q)$ can be obtained approximately from the formula

$$ {}_\nu P_k(Q) \doteq \{{}_\nu P_1(Q)\}^k. \tag{4} $$

* This appears as $1 - M$ in equation (10) of Finney's paper through a printing error.

For $m = 2$, this becomes $$ {}_{\nu}P_k(Q) \simeq \left\{1 - \left(1 + \frac{2Q^2}{\nu}\right)^{-\frac{1}{2}\nu}\right\}^k. \tag{5}$$

To get some idea of the adequacy of the approximate approach, Finney compared the 5 % points for F_k obtained by the two methods, when $m = 2$. He concluded that the approximation was satisfactory when ν was moderately large in comparison with k, and presumed that the agreement would generally improve with increase of m. The chief uncertainty according to him was, therefore, when $m = 1$.

An alternative method of calculating (1) which is particularly suitable when $m = 1$ is Hartley's expansion of a studentized integral. Thus, using (1) of Part I, the integral itself is

$$ {}_{\nu}P_k(Q) = 2\Gamma(\tfrac{1}{2}\nu)^{-1}(\tfrac{1}{2}\nu)^{\frac{1}{2}\nu}\int_0^\infty s_0^{\nu-1}e^{-\frac{1}{2}\nu s_0^2} P_k(Qs_0)\, ds_0, \tag{6}$$

where $$ P_k(Q) = \left(2\Gamma(\tfrac{1}{2}m)^{-1}(\tfrac{1}{2}m)^{\frac{1}{2}m}\int_0^Q x^{m-1}e^{-\frac{1}{2}mx^2}\, dx\right)^k. \tag{7}$$

It could easily be verified that (6) and (7) lead to the same result as given in (3) when $m = 2$.

In the form given by (6), ${}_{\nu}P_k(Q)$ can be evaluated with sufficient accuracy using expansion (6) of Part I where $P(Q)$ should be replaced by $P_k(Q)$ given in (7). The coefficients $a_0, a_1, a_2, \ldots$, of this expansion will be seen to involve powers and derivatives of the incomplete gamma-function, thus leading to Laguerre functions. When $m = 1$ or 2 these reduce to certain Hermite functions which are fully tabulated in the *British Association Mathematical Tables*, vol. I. In the general case, the Laguerre functions have to be used; these have been tabulated, although less extensively.

In the special case where $m = 1$ we have

$$ P_k(Q) = \left(\sqrt{\frac{2}{\pi}}\int_0^Q e^{-\frac{1}{2}x^2}\, dx\right)^k. \tag{8}$$

This can be expressed in terms of the Hermite function

$$ Hh_0(x) = \int_x^\infty e^{-\frac{1}{2}x^2}\, dx \tag{9}$$

tabulated in the *British Association Mathematical Tables*, which also give values of

$$ Hh_{-n}(x) = \left(-\frac{d}{dx}\right)^n Hh_0(x) \tag{10}$$

for $n = 1$ to 7. These have been used in preparing the tables described in § 3.

When $m = 2$, we have

$$ \begin{aligned} P_k(Q) &= (1 - e^{-Q^2})^k \\ &= \{1 - Hh_{-1}(x)\}^k, \quad \text{where} \quad x = Q\sqrt{2}. \end{aligned} \tag{11}$$

The first six derivatives of the right-hand side of (11) can be obtained from the *British Association Mathematical Tables*. We could therefore proceed with Hartley's method for the case $m = 2$ as well. It seems simpler, however, to use the exact formula (3). Finney has prepared a table of percentage points when $k = 2$ and 3. We shall consider only the case $m = 1$.

3. *Construction of tables*

Writing $P_k(Q)$ as P_k for brevity, we have, for $m = 1$,

$$\left.\begin{aligned}
(\tfrac{1}{2}\pi)^{\frac{1}{2}k} P_k &= \{\sqrt{\tfrac{1}{2}\pi} - Hh_0(Q)\}^k = A^k \text{ (say)},\\
(\tfrac{1}{2}\pi)^{\frac{1}{2}k} P_k' &= kA^{k-1}Hh_{-1},\\
(\tfrac{1}{2}\pi)^{\frac{1}{2}k} P_k'' &= k(k-1)A^{k-2}Hh_{-1}^2 - kA^{k-1}Hh_{-2},\\
(\tfrac{1}{2}\pi)^{\frac{1}{2}k} P_k''' &= k(k-1)(k-2)A^{k-3}Hh_{-1}^3 - 3k(k-1)A^{k-2}Hh_{-1}Hh_{-2} + kA^{k-1}Hh_{-3},\\
(\tfrac{1}{2}\pi)^{\frac{1}{2}k} P_k^{\text{iv}} &= k(k-1)(k-2)(k-3)A^{k-4}Hh_{-1}^4 - 6k(k-1)(k-2)A^{k-3}Hh_{-1}^2 Hh_{-2}\\
&\quad + k(k-1)A^{k-2}(3Hh_{-2}^2 + 4Hh_{-1}Hh_{-3}) - kA^{k-1}Hh_{-4},\\
(\tfrac{1}{2}\pi)^{\frac{1}{2}k} P_k^{\text{v}} &= k(k-1)(k-2)(k-3)(k-4)A^{k-5}Hh_{-1}^5 - 10k(k-1)(k-2)(k-3)A^{k-4}Hh_{-1}^3 Hh_{-2}\\
&\quad + 5k(k-1)(k-2)A^{k-3}Hh_{-1}(3Hh_{-2}^2 + 2Hh_{-1}Hh_{-3})\\
&\quad - 5k(k-1)A^{k-2}(2Hh_{-2}Hh_{-3} + Hh_{-1}Hh_{-4}) + kA^{k-1}Hh_{-5},\\
(\tfrac{1}{2}\pi)^{\frac{1}{2}k} P_k^{\text{vi}} &= k(k-1)(k-2)(k-3)(k-4)(k-5)A^{k-6}Hh_{-1}^6\\
&\quad - 15k(k-1)(k-2)(k-3)(k-4)A^{k-5}Hh_{-1}^4 Hh_{-2}\\
&\quad + 5k(k-1)(k-2)(k-3)A^{k-4}(9Hh_{-1}^2 Hh_{-2}^2 + 4Hh_{-1}^3 Hh_{-3})\\
&\quad - 15k(k-1)(k-2)A^{k-3}(Hh_{-2}^3 + 4Hh_{-1}Hh_{-2}Hh_{-3} + Hh_{-1}^2 Hh_{-4})\\
&\quad + k(k-1)A^{k-2}(10Hh_{-3}^2 + 15Hh_{-2}Hh_{-4} + 6Hh_{-1}Hh_{-5}) - kA^{k-1}Hh_{-6}.
\end{aligned}\right\} \tag{12}$$

These six derivatives are sufficient to calculate a_0, a_1, a_2 and a_3 of expansion (6) of Part I. Since a_4 involves Hh_{-8} which is not included in the *British Association Mathematical Tables*, it is not easy to calculate a_4 by the direct method. The calculation of a_3 itself is rather complicated by direct evaluation from the derivatives of P_k and was therefore accomplished by an indirect method of numerical differentiation from a_1 and a_2. The same method could have been used to calculate a_4, but the idea was abandoned as that term does not appear seriously to affect the accuracy, unless k is large and ν is small. A small panel of values of a_4 was, however, calculated by this method for select combinations of values of Q and k to get some idea of its range of magnitude. These are discussed later in this section.

Since values of $Q^r P_k^{(r)}$ were required, an auxiliary function

$$h_i = Hh_{-i}Q^i A^{i-1} \tag{13}$$

was introduced in terms of which $Q^r P_k^{(r)}$ became, for $r = 1, 2, 3$ and 4,

$$\left.\begin{aligned}
(\tfrac{1}{2}\pi)^{\frac{1}{2}k} QP' &= kA^{k-1}h_1,\\
(\tfrac{1}{2}\pi)^{\frac{1}{2}k} Q^2P'' &= kA^{k-2}\{(k-1)h_1^2 - h_2\},\\
(\tfrac{1}{2}\pi)^{\frac{1}{2}k} Q^3P''' &= kA^{k-3}\{(k-1)(k-2)h_1^3 - 3(k-1)h_1h_2 + h_3\}\\
(\tfrac{1}{2}\pi)^{\frac{1}{2}k} Q^4P^{\text{iv}} &= kA^{k-4}\{(k-1)(k-2)(k-3)h_1^4 - 6(k-1)(k-2)h_1^2h_2\\
&\qquad + (k-1)(3h_2^2 + 4h_1h_3) - h_4\}.
\end{aligned}\right\} \tag{14}$$

Expressions for a_0, a_1 and a_2 are

$$\left.\begin{aligned}
a_0 &= \left(\sqrt{\frac{2}{\pi}}A\right)^k,\\
a_1 &= \tfrac{1}{4}k(\tfrac{1}{2}\pi)^{-\frac{1}{2}k}[\{(k-1)h_1^2 - h_2\}A^{k-2} - h_1A^{k-1}],\\
a_2 &= \tfrac{1}{96}(3Q^4P^{\text{iv}} - 2Q^3P''') - \tfrac{1}{8}a_1.
\end{aligned}\right\} \tag{15}$$

It was decided to limit the values of k to $\leqslant 10$. Using the *British Association Mathematical Tables*, values of A^k, h_1, h_2, h_3 and h_4 were calculated to six decimal accuracy in the range 1·6 (0·2) 6·6 for Q. Values of a_0, a_1 and a_2 were then obtained and all the calculations checked by fifth order differencing on the National Accounting Machine.

The method used to obtain a_3 was to express it in terms of a_1, a_2 and their derivatives a_1', a_2' and a_2''. Thus

$$a_3 = \tfrac{1}{12}Q^2a_2'' - \tfrac{19}{36}Qa_2' + \tfrac{16}{27}a_2 + \tfrac{4}{81}(Qa_1' - 4a_1). \tag{16}$$

The values of a_1', a_2' and a_2'' were obtained by numerical differentiation, which gave the final values of a_3 to two decimal accuracy; this was ample for our purpose. An independent check was provided for a few test values, taking $k = 2$. For obtaining these values, P^{iv}, P^{v} and P^{vi} were calculated and substituted in the formula

$$a_3 = \tfrac{1}{384}(Q^6P^{\text{vi}} + Q^5P^{\text{v}} + 11Q^4P^{\text{iv}}) - \tfrac{3}{2}a_2 - \tfrac{11}{32}a_1 \tag{17}$$

and compared with those obtained with the help of (16). The comparison is set out in Table 1. The agreement is satisfactory for our purpose.

Table 1. *Values of* a_3

Q	Formula (16)	Formula (17)
2·2	1·8190	1·8242
2·4	2·3353	2·3350
2·6	2·3178	2·3176
2·8	1·7453	1·7449
3·0	0·7801	0·7799

A further short-cut was used in calculating a_3. Instead of using (16) for each value of k ranging from 2 to 10, a_3 was first calculated for $k = 2, 6$ and 10. By three-point interpolation between them, values of a_3 were obtained for the remaining values of k.

Although a_4 was not used in the evaluation of ${}_\nu P_k(Q)$, a few of its values were calculated by numerical differentiation of a_1 and a_2, using the formula

$$a_4 = \tfrac{1}{192}Q^4a_2^{\text{iv}} - \tfrac{17}{288}Q^3a_2''' + \tfrac{439}{1728}Q^2a_2'' - \tfrac{9401}{25920}Qa_2' - \tfrac{62}{243}a_2 + \tfrac{44}{3645}(Qa_1' - 4a_1). \tag{18}$$

These values are presented in Table 2.

Table 2. *Values of* a_4

Q \ k	2	6	10
2·8	+ 2·5	+ 0·1	− 60·0
3·6	+12·0	+50·0	+108·9
4·4	− 6·1	−18·1	− 28·1

The contribution from a_4 to the probability integral may be as large as 0·01 near the upper 5 and 1 % points if $k = 10$, $\nu = 10$. This will affect the percentage point of the largest variance ratio by as much as 1 in the unit's place. If $\nu = 20$, only the first decimal will be affected by about 1 or 2. It is likely in practice that ν will be of the order of 20 or more when $k = 10$, so that the table will not vitiate the level of significance to any serious extent.

The values of ν selected for constructing the tables of the upper 5 and 1 % points of F_k were 10, 12, 15, 20, 30, 60 and ∞ to facilitate harmonic interpolation inside this range. Second-difference inverse interpolation was used to obtain the percentage points from the table of probability integrals.

Table 3 shows the final results giving the upper 5 and 1 % points for $k = 1$ to 10. The values for $k = 1$ were copied from Merrington & Thompson's (1943) tables, and the difference between these and the kth column ($k > 1$) shows how much out we might be in assessing the significance of each variance ratio singly, rather than as one value in a group of k. Each row and column of the percentage points in Table 3 were differenced to the third order as a final check. No suspicious-looking values were found.

Table 3. *Upper per cent points of the largest variance ratio**

5 % points

ν \ k	1	2	3	4	5	6	7	8	9	10
10	4·96	6·79	8·00	8·96	9·78	10·52	11·18	11·79	12·36	12·87
12	4·75	6·44	7·53	8·37	9·06	9·68	10·20	10·68	11·12	11·53
15	4·54	6·12	7·11	7·86	8·47	8·98	9·43	9·82	10·19	10·52
20	4·35	5·81	6·72	7·40	7·94	8·39	8·79	9·13	9·44	9·71
30	4·17	5·52	6·36	6·97	7·46	7·87	8·21	8·51	8·79	9·03
60	4·00	5·25	6·02	6·58	7·02	7·38	7·68	7·96	8·20	8·41
∞	3·84	5·00	5·70	6·21	6·60	6·92	7·20	7·44	7·65	7·84

1 % points

ν \ k	1	2	3	4	5	6	7	8	9	10
10	10·04	13·17	15·08	16·43	17·43	18·25	18·91	19·48	19·97	20·41
12	9·33	11·88	13·52	14·73	15·69	16·47	17·12	17·68	18·16	18·60
15	8·68	10·82	12·18	13·21	14·03	14·72	15·30	15·81	16·26	16·66
20	8·10	9·93	11·08	11·93	12·61	13·19	13·67	14·09	14·49	14·83
30	7·56	9·16	10·14	10·86	11·43	11·90	12·31	12·66	12·97	13·26
60	7·08	8·49	9·34	9·95	10·43	10·82	11·15	11·45	11·72	11·95
∞	6·63	7·88	8·61	9·15	9·54	9·87	10·16	10·41	10·62	10·82

* k is the number of independent variance estimates, each based on 1 degree of freedom. ν denotes the degrees of freedom of the independent 'error' variance.

4. *An illustration*

A good example to illustrate a possible application of Table 3 is provided in Wishart's (1938) paper. He analysed the data (number of sticks) of a uniformity trial on asparagus as if it were an experiment with nitrate (N), phosphate (P) and potash (K) at two levels each, in eight blocks of four plots confounding the three-factor interaction between blocks.† The analysis of variance is set out in Table 4.

† The purpose of this preliminary trial was to collect material for adjustment, by the covariance technique, of inherent fertility differences among the plots when the N, P, K treatments were given in the ensuing season.

Wishart writes: 'The mean square for total "treatments" (6 D.F.) is not significant, but on examination of the separate effects it would appear that the PK interaction is significant at the 5 % level [$F_{0\cdot05} = 4\cdot41$, as against the observed $F = 7\cdot02$.] With many ordinary experiments we should be happy to claim such a result as indicating a real effect.... But since the treatments were not applied, the effect is entirely accidental.'

He then applies Bartlett's test for homogeneity of variance (which in this case is equivalent to the Neyman-Pearson L_1 test) on the six treatment variances and finds that 'the set of mean squares is compatible with the hypothesis that it is homogeneous, and awkward explanations are avoided'.

D. J. Bishop & U. S. Nair (1939) found that Bartlett's test underestimates the significance if the degrees of freedom of some of the k variances are as small as 1 or 2. The 5 % point of L_1 which they give when $k = 6$ and $\nu = 1$ is 0·094 against Bartlett's value of 0·077. The value of L_1 for the present example is 0·152. It is not significant,* even by the exact L_1-test of Bishop & Nair.

Table 4. *Analysis of variance*

Variation	Degrees of freedom	Mean square (variance)	Variance ratio (F)
Blocks	7	6543·2	—
N	1	488·3	0·35
P	1	69·0	0·05
K	1	34·0	0·02
NP	1	830·3	0·60
NK	1	57·8	0·04
PK	1	9765·0	7·02
(Total treatments)	(6)	(1874·1)	—
Error	18	1391·2	—

Since only *one* of the six variances is standing out prominently, Cochran's (1941) test based on g, the ratio of the largest variance divided by the sum of the set of variances may appeal as more appropriate than the L_1-test. The value of g for the six treatment variances of Table 4 is 0·87. The 5 % value of g given in Cochran's table is 0·78. According to the g-test, therefore, the variance for interaction PK is significantly large, making us reject a hypothesis, which we know to be true.

To decide which of the two tests will be more appropriate in a situation where the heterogeneity among the k variances is caused by the presence of a single variance much larger than the rest, would need a study of the power functions of the two tests for this particular class of alternative hypotheses, using the Neyman-Pearson theory.

In the present example, however, neither the L_1 nor the g-test takes into account the estimated 'error' variance with the help of which it has been possible to say that the total treatment effects (6 D.F.) were *not* significant. The question now to be answered is whether the variance due to interaction PK, being the largest of the six individual treatment effects, could be considered as significantly greater than the error variance. The variance ratio for PK is 7·02. The 5 % point in Table 3 for $k = 6$, $\nu = 18$ is 8·59. The observed largest variance ratio is, therefore, not significantly large at the 5 % level.

* For L_1, significance is indicated by *low*, not high, values.

5. *Probability integral of the smallest variance ratio*

Let ${}_\nu p_k(q)$ denote the probability that s_1/s_0 is $\geqslant q$. The probability that the smallest variance ratio F_1 is $\geqslant q^2$ will also be given by ${}_\nu p_k(q)$.

Finney (1941) showed that

$$ {}_\nu p_k(q) = (1-M)^k (1+\lambda)^{-\frac{1}{2}\nu}, \tag{19} $$

where M is as defined in (2) except for a replacement of Q by q; and λ is to be put equal to zero after differentiation.

Using Hartley's general studentized integral

$$ {}_\nu p_k(q) = 2\Gamma(\tfrac{1}{2}\nu)^{-1} (\tfrac{1}{2}\nu)^{\frac{1}{2}\nu} \int_0^\infty s_0^{\nu-1} e^{-\frac{1}{2}\nu s_0^2} p_k(qs_0)\, ds_0, \tag{20} $$

where

$$ p_k(q) = \left(2\Gamma(\tfrac{1}{2}m)^{-1} (\tfrac{1}{2}m)^{\frac{1}{2}m} \int_q^\infty x^{m-1} e^{-\frac{1}{2}mx^2} dx\right)^k. \tag{21} $$

Finney remarks: 'The convergence of the series by which significance levels for the smallest variance ratio are obtained is much more rapid than for the largest ratio. Also, as Hartley has demonstrated, the approximation given by an assumption of independence is very close even for small numbers of degrees of freedom.'

These remarks could easily be substantiated when $m = 2$ for which the exact value of ${}_\nu p_k(q)$ obtained from either (19) or (20) is

$$ \left(1+\frac{2kq^2}{\nu}\right)^{-\frac{1}{2}\nu} \tag{22} $$

On the assumption of independence as in (4), the approximate value for ${}_\nu p_k(q)$ will be

$$ \left(1+\frac{2q^2}{\nu}\right)^{-k\nu} \tag{23} $$

Evidently (22) and (23) are not identical,* but when q is small and ν is not too small, they will be very close to each other.

As we are concerned with the lower percentage points of the smallest variance ratio, q will be small and hence approximation (4) should yield sufficiently accurate results for any m.

We may, however, examine this more closely for the case $m = 1$. The exact value of ${}_\nu p_k(q)$ can be obtained using the expansion (6) of Part I. Owing to the smallness of q, it is scarcely necessary to add terms beyond a_1. We may, however, include a_2 which is approximately equal to $-a_1/8$. The a_0 and a_1 terms are calculated from the formulae

$$ a_0 = p_k(q) = \left(\sqrt{\frac{2}{\pi}} \int_q^\infty e^{-\frac{1}{2}x^2} dx\right)^k = (1-\alpha)^k, \tag{24} $$

$$ a_1 = k(qz)\,(1-\alpha)^{k-2} \{(k-1)\,qz + \tfrac{1}{2}(q^2+1)\,(1-\alpha)\}, \tag{25} $$

where α and z are quantities given in Table 2 of the *Tables for Statisticians and Biometricians*, Part I.

In Table 5, values of a_0 and a_1 are calculated for q in the range 0 to 0·1 at intervals of 0·01 and for $k = 1$ to 10. The probability that the smallest of k variance ratios, F_1 is $\leqslant q^2$ will be given by

$$ 1 - {}_\nu p_k(q) = \left\{1 - a_0 - \frac{a_1}{\nu}\left(1 - \frac{1}{8\nu}\right)\right\}, \tag{26} $$

* Finney claims in equation (21) of his paper that they are identical, obviously by oversight.

Table 5. *Table for calculating the probability integral of the smallest variance ratio**

q \ k	2		3		4	
	a_0	a_1	a_0	a_1	a_0	a_1
0·00	1·000 000	0·000 00	1·000 000	0·000 00	1·000 000	0·000 00
0·01	0·984 106	0·003 99	0·976 254	0·005 98	0·968 465	0·007 98
0·02	0·968 341	0·007 98	0·952 890	0·011 97	0·937 685	0·015 95
0·03	0·952 707	0·011 97	0·929 906	0·017 95	0·907 650	0·023 90
0·04	0·937 204	0·015 97	0·907 301	0·023 93	0·878 352	0·031 84
0·05	0·921 835	0·019 97	0·885 074	0·029 90	0·849 780	0·039 74
0·06	0·906 600	0·023 97	0·863 225	0·035 87	0·821 924	0·047 61
0·07	0·891 502	0·027 98	0·841 750	0·041 83	0·794 775	0·055 43
0·08	0·876 540	0·032 00	0·820 649	0·047 78	0·768 323	0·063 20
0·09	0·861 717	0·036 02	0·799 921	0·053 72	0·742 556	0·070 90
0·10	0·847 034	0·040 05	0·779 563	0·059 64	0·717 466	0·078 53

q \ k	5		6		7	
	a_0	a_1	a_0	a_1	a_0	a_1
0·00	1·000 000	0·000 00	1·000 000	0·000 00	1·000 000	0·000 00
0·01	0·960 738	0·009 97	0·953 072	0·011 96	0·945 468	0·013 95
0·02	0·922 723	0·019 92	0·907 999	0·023 88	0·893 511	0·027 83
0·03	0·885 927	0·029 83	0·864 725	0·035 72	0·844 029	0·041 56
0·04	0·850 327	0·039 68	0·823 196	0·047 44	0·796 930	0·055 09
0·05	0·815 893	0·049 46	0·783 357	0·059 00	0·752 118	0·068 36
0·06	0·782 600	0·059 13	0·745 157	0·070 38	0·709 505	0·081 30
0·07	0·750 421	0·068 69	0·708 543	0·081 53	0·669 002	0·093 88
0·08	0·719 332	0·078 11	0·673 465	0·092 42	0·630 523	0·106 05
0·09	0·689 306	0·087 38	0·639 874	0·103 03	0·593 986	0·117 76
0·10	0·660 316	0·096 48	0·607 718	0·113 34	0·559 310	0·128 98

q \ k	8		9		10	
	a_0	a_1	a_0	a_1	a_0	a_1
0·00	1·000 000	0·000 00	1·000 000	0·000 00	1·000 000	0·000 00
0·01	0·937 924	0·015 94	0·930 440	0·017 92	0·923 017	0·019 90
0·02	0·879 253	0·031 76	0·865 223	0·035 67	0·851 417	0·039 56
0·03	0·823 829	0·047 35	0·804 112	0·053 09	0·784 867	0·058 75
0·04	0·771 503	0·062 63	0·746 886	0·070 03	0·723 056	0·077 29
0·05	0·722 126	0·077 50	0·693 329	0·086 39	0·665 681	0·095 03
0·06	0·675 559	0·091 88	0·643 237	0·102 06	0·612 462	0·111 83
0·07	0·631 667	0·105 70	0·596 416	0·116 95	0·563 132	0·127 60
0·08	0·590 320	0·118 92	0·552 679	0·131 00	0·517 439	0·142 25
0·09	0·551 390	0·131 47	0·511 848	0·144 14	0·475 142	0·155 72
0·10	0·514 758	0·143 32	0·473 755	0·156 32	0·436 017	0·167 97

* $1-{}_{\nu}p_k(q) = 1-a_0-\frac{a_1}{\nu}\left(1-\frac{1}{8\nu}\right)$, see equation (26).

which can be calculated with the help of Table 5 for any given q, k and ν. If this probability is less than 0·05 or 0·01, the observed smallest variance ratio q^2 can be said to be significantly small at the 5 or 1 % level respectively.

The values of ${}_\nu p_k(q)$ given by Table 5 have been compared in Table 6 with those obtained by the approximate formula (4), for $\nu = 10$, $q = 0{\cdot}10$ and 0·01.

Table 6. *Comparison of exact and approximate values of the probability integrals* of the smallest variance ratio*

k	$q = 0{\cdot}10$		$q = 0{\cdot}01$	
	${}_{10}p_k(q)$	$\{{}_{10}p_1(q)\}^k$	${}_{10}p_k(q)$	$\{{}_{10}p_1(q)\}^k$
1	0·922 324	0·922 321	0·992 218	0·992 218
2	0·850 989	0·850 676	0·984 500	0·984 497
3	0·785 452	0·784 596	0·976 845	0·976 836
4	0·725 221	0·723 649	0·969 253	0·969 234
5	0·669 843	0·667 437	0·961 722	0·961 691
6	0·618 911	0·615 591	0·954 253	0·954 207
7	0·572 047	0·567 773	0·946 845	0·946 781
8	0·528 911	0·523 669	0·939 498	0·939 413
9	0·489 191	0·482 990	0·932 210	0·932 102
10	0·452 604	0·445 472	0·924 982	0·924 848

There is satisfactory agreement between the two methods. Indeed, as is to be expected, the agreement is much better for $q = 0{\cdot}01$ than for $q = 0{\cdot}10$. The difference in the top row values in columns (2) and (3) is due to calculating the latter up to the term in ν^{-4} using the incomplete beta-function expansion (10) of Part I. It will be seen that for this expansion, terms up to ν^{-2} give a five decimal accuracy when $q = 0{\cdot}10$ and this improves to six decimals when $q = 0{\cdot}01$.

We may, before concluding, add an illustration for the use of the test of the smallest variance ratio. This is also taken from Wishart's (1938) paper, where by the ordinary test of significance he gets variance for an interaction NP as significantly less than the error variance. He calls this a '*negative* interaction'. Applying Bartlett's test, Wishart found that this 'negative' interaction was not significant. The values concerned are $s_1^2 = 0{\cdot}01$, $s_0^2 = 5{\cdot}57$, $k = 6$, $m = 1$, $\nu = 18$. To apply the test for significance of the smallest variance ratio, we calculate $F_1 = \dfrac{0{\cdot}01}{5{\cdot}57} = 0{\cdot}0018$ giving $q = 0{\cdot}04$. Using (26) and referring to Table 5, we get

$$ {}_{18}p_6(0{\cdot}04) = a_0 + a_1(1 - 8/\nu)/\nu = 0{\cdot}8258. $$

The probability that F_1 is less than or equal to the observed value is about 0·17, which is much greater than 0·05. The observed F_1 is therefore not significantly small.

In conclusion, I should like to acknowledge warmly the help and guidance I have received from Prof. E. S. Pearson and Dr H. O. Hartley in the course of my investigations.

* The probability integral is calculated here from q to ∞, instead of the usual limits 0 to q.

REFERENCES

BISHOP, D. J. & NAIR, U. S. (1939). A note on certain methods of testing for the homogeneity of a set of estimated variances. *J. Roy. Statist. Soc. Suppl.* **6**, 89.

COCHRAN, W. G. (1941). The distribution of the largest of a set of estimated variances as a fraction of their total. *Ann. Eugen., Lond.*, **11**, 47.

FINNEY, D. J. (1941). The joint distribution of variance ratios based on a common error mean square. *Ann. Eugen., Lond.*, **11**, 136.

FISHER, R. A. (1924). On a distribution yielding the error functions of several well-known statistics. *Proc. Intern. Math. Congress*, Toronto, pp. 805–13.

FISHER, R. A. (1926). Expansion of 'Student's' integral in powers of n^{-1}. *Metron*, **5**, 109.

HARTLEY, H. O. (1938). Studentization and large-sample theory. *J. Roy. Statist. Soc. Suppl.* **5**, 80.

HARTLEY, H. O. (1944). Studentization, or the elimination of the standard deviation of the parent population from the random sample-distribution of statistics. *Biometrika*, **33**, 173.

MAHALANOBIS, P. C. (1944). On large-scale sample surveys. *Phil. Trans. Roy. Soc.* B, **231**, 329.

MERRINGTON, M. & THOMPSON, C. M. (1943). Tables of percentage points of the inverted Beta (F) distribution. *Biometrika*, **33**, 73.

NAIR, K. R. (1948). The distribution of the extreme deviate from the sample mean and its studentized form. *Biometrika*, **35**, 118.

NEYMAN, J. & PEARSON, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, **20 A**, 175, 263.

PEARSON, E. S. & HARTLEY, H. O. (1943). Tables of the probability integral of the 'studentized' range. *Biometrika*, **33**, 89.

'STUDENT' (1908). On the probable error of a mean. *Biometrika*, **6**, 1.

SUKHATME, P. V. (1937). Tests of significance for samples of the χ^2-population with two degrees of freedom. *Ann. Eugen., Lond.*, **8**, 52.

THOMPSON, C. M. (1941). Tables of percentage points of the incomplete beta-function. *Biometrika*, **32**, 151.

WISHART, J. (1927). On the approximate quadrature of certain skew curves with an account of the researches of Thomas Bayes. *Biometrika*, **19**, 1.

WISHART, J. (1938). Field experiments of factorial design. *J. Agric. Sci.* **28**, 299.

THE ESTIMATION OF NON-LINEAR PARAMETERS BY 'INTERNAL LEAST SQUARES'

By H. O. HARTLEY

1. Introduction

(1·1) *The difficulties arising with curvilinear regressions*

In statistical regression of two variables y and x, the linear relation $y = a + bx$ has received most attention. Apart from occasional 'fitting of polynomials', non-linear relations have not been used frequently. As reasons for the reluctance on the part of statistical workers to use non-linear regressions we may mention here:

(*a*) The complication in the computational procedure when estimating non-linear parameters by efficient methods such as least squares.

(*b*) The lack of exactness in tests of the goodness of fit and the difficulty of establishing the random sampling distribution of fitted statistics.

(*c*) The fact that any transformations to linearity such as $f(y) = bg(x) + a$ (which may be suggested by theory), usually involve certain unknown parameters in f and g which must be estimated from the sample and, therefore, bring in difficulties of type (*b*).

(*d*) The difficulty of deciding which of the many possible non-linear regressions is suggested by theory.

As examples illustrating these difficulties we may mention here:

(*a*) The iterative process in dosage mortality technique which necessitates the preparation of special auxiliary tables for each fitted law.

(*b*) Methods of fitting the Gompertz curve by factorial moments (Sasuly, 1934) resulting in procedures of unknown efficiency.

(*c*) The exponential law of diminishing returns:

$$y = \hat{y}(1 - e^{kx}) \tag{1}$$

can be transformed to a linear relation between $\log_e(1 - y/\hat{y})$ and x:

$$\log_e(1 - y/\hat{y}) = kx. \tag{1'}$$

However, a knowledge of the 'limiting response' $\hat{y}$ is usually required in order to make this transformation to linearity. If $\hat{y}$ is assessed by 'inspection' the accuracy of the subsequent fit of k cannot be ascertained, let alone the accuracy of the assessment of $\hat{y}$.

(*d*) In numerous applications polynomial regressions have been fitted without a biological or social theory to guide the fit and, as a result, difficulties have been encountered in interpreting the real meaning of the polynomial terms.

(1·2) *The principle of 'internal regression'*

The principle here adopted is that of bringing most of the important regression laws under one generating principle: Most of the laws in physics and technology are generated by simple, mainly linear, relationships, between the function and its first and second derivative (e.g. velocity and acceleration) and the same applies to most of the important laws in the biological sciences. We may name two examples:

(A) The exponential law mentioned above is generated by the first order differential equation

$$\frac{dy}{dx} = -k\hat{y} + ky, \tag{2}$$

which is a linear relationship. If applied to fertilizer response in agriculture ($k < 0$), this has the simple meaning that the additional yield, dy, caused by an addition of fertilizer dx, is, in the first place, proportional to dx (i.e. $= \hat{y}(-k)\,dx$); but this constant rate of increase ($\hat{y}(-k)$) is retarded by an amount $(-(-k)y)$ which in turn is proportional to the yield y already attained.

(B) The 'trade cycle' for, say, a production index y is sometimes regarded as being caused by the interplay of production y and demand z which are assumed to satisfy the relations

$$\frac{dy}{dt} = \alpha z, \quad \frac{dz}{dt} = -\beta y,$$

resulting in

$$\frac{d^2y}{dt^2} = -\alpha\beta y \tag{3}$$

of which the general solution is the trade cycle law:

$$y = \sin[\sqrt{(\alpha\beta)}\,t + \gamma]. \tag{4}$$

The idea, now, is to fit directly to the data the generating *linear* law (such as (2) or (3)) resulting in linear equations for the parameter estimates, rather than fitting the non-linear regressions (such as (1) or (4)). However, some modications are required: whilst for a mathematical function $y(x)$ the regression (1) and the differential equation (2) are identical conditions, it would be difficult to fit to an empirical series of observed y_i a condition involving its differential coefficient. We therefore seek a finite difference equivalent to the differential equation (2). As is well known (see, for example, Bartlett, 1946; Cunningham & Hynd, 1946; Kendall, 1944) the first order linear difference equation of the form

$$y_i - y_{i-1} = by_i + a \tag{5}$$

is capable of generating exactly the exponential law of type (1). Integrating (or summing) (5) we obtain

$$y_j = bY_j + ax_j + c, \tag{6}$$

where $Y_j = \sum_{i=0}^{j} y_i$. Now this equation may be regarded as a linear regression equation for the dependent variable y_j with its own progressive sum Y_j as independent variable, and with x_j playing the part of a second independent variable. Formally, therefore, the non-linear parameters $\hat{y}$ and k of (1) can be estimated from the simple linear regression coefficients a, b, c given in (6).†

In a similar manner the second order differential equation (3) can be replaced by a linear relation between y_j and its second sum, ${}_2Y_j = \sum_{l=0}^{j}\sum_{i=0}^{l} y_i$, and such a relation can be shown to be equivalent to the harmonic regression law (trade cycle) (4). The period $\sqrt{(\alpha\beta)}$ of this trade cycle can therefore be determined as a partial linear regression coefficient between y_j and ${}_2Y_j$.

Turning now to the general case, we shall call any regression equation in which the dependent variate y is related to its own repeated sums

$${}_1Y_j = \sum_{i=0}^{j} y_i, \quad {}_2Y_j = \sum_{l=0}^{j}\sum_{i=0}^{l} y_i \tag{7}$$

† In §(2·1) we shall give the exact relation between the parameters y_0, k and a, b, c.

etc., as independent variables, an 'Internal regression'. This concept is, of course, closely linked with that of auto-regression in time series. Its usefulness lies in its close relation to linear differential equations, which are known to generate most of the important regression laws and also give a better understanding of the physical or biological mechanism producing the respective curvilinear regression.

The principle of estimation and goodness of fit that we shall use in conjunction with internal regression is that of least squares applied to the 'integrated' equation (6) and not to the corresponding difference equation (5) (see, for example, Bartlett, 1946; Mann & Wald, 1943). This method is, at first, accepted as a working rule and the formulae for the estimators derived and demonstrated in terms of examples. Only the first order internal regression is treated here (§ 2). The efficiency of the estimator is then considered under various assumptions of residual independence. In particular, the classical assumption of random residuals attached to the observed y_i is fully investigated by comparing the efficiency of our estimators with the 100 % efficient transcendental maximum likelihood roots. To this end, large sample formulae for the variances of the estimators will be developed (§ 3), leaving the derivation of exact sampling distributions for goodness of fit tests to a later paper.

2. The procedure of internal regression

(2·1) *The first order internal regression; fitting of the exponential law*

Consider a sample of observed values y_i corresponding to discrete integer positions, $x_i = i$, of the independent variable x (e.g. a time series). If the y_i satisfy exactly the difference equation

$$(y_{i+1} - y_i) = -\tfrac{1}{2}b(y_{i+1} + y_i) + a \tag{8}$$

then, by standard finite difference technique, they also satisfy exactly the exponential law

$$y_i = \hat{y}(1 - fe^{kx_i}), \tag{9}$$

where the limiting response $\hat{y}$ and the 'exponential curvature' k are given by

$$\hat{y} = a/b; \quad -k = 2\tanh^{-1}\tfrac{1}{2}b \tag{10}$$

whilst f is a constant of integration. In order to find the least square estimates for a and b it is convenient to distinguish two cases:

(2·11) $n = \text{odd} = 2m+1 \quad (i = -m, -m+1, \ldots, -1, 0, 1, \ldots, m)$,

(2·12) $n = \text{even} = 2m \quad (i = -m, \ldots, -1, 1, \ldots, m)$.

(2·11) We introduce

$$\bar{y} = (2m+1)^{-1} \sum_{i=-m}^{+m} y_i, \quad \eta_i = y_i - \bar{y}, \quad \xi_i = i. \tag{11}$$

Hence

$$\Sigma\eta_i = \Sigma\xi_i = 0. \tag{12}$$

We rewrite (8) as

$$(\eta_{i+1} - \eta_i) = -\tfrac{1}{2}b(\eta_{i+1} + \eta_i) + (a - b\bar{y}) \tag{13}$$

and sum the difference equation form $i = 0$ to $i = j$, introducing at the same time as a new variable S_j the first sum of the η_i as follows:

$$\left.\begin{aligned} S_j &= \tfrac{1}{2}\eta_0 + \sum_{i=1}^{j-1}\eta_i + \tfrac{1}{2}\eta_j \quad \text{for} \quad j \geqslant 1, \\ S_0 &= 0, \\ S_j &= -\tfrac{1}{2}\eta_0 - \sum_{i=-1}^{j+1}\eta_i - \tfrac{1}{2}\eta_j \quad \text{for} \quad j \leqslant -1. \end{aligned}\right\} \tag{14}$$

The progressive sums of (13) are then found to be equivalent to the equations:

$$\eta_j - \eta_0 = -bS_j + (a - b\bar{y})\,\xi_j + c, \tag{15}$$

or

$$\eta_j = -bS_j + a'\xi_j + c', \tag{16}$$

where

$$a' = a - b\bar{y},$$

or

$$a'/b = y_0 - \bar{y} \tag{17}$$

and c, c' are constants of summation.

The least square solutions for a', $-b$ and c' are then given by:

$$\left.\begin{aligned} -b\Sigma S_i^2 \;+a'\Sigma S_i\xi_i + c'\Sigma S_i &= \Sigma S_i\eta_i, \\ -b\Sigma S_i\xi_i + a'\Sigma\xi_i^2 \qquad\quad &= \Sigma\eta_i\xi_i, \\ -b\Sigma S_i \qquad\quad +c'n &= 0. \end{aligned}\right\} \tag{18}$$

It is now easy to see from the definition of the S_j that they are exactly orthogonal to the η_j, i.e. that

$$\sum_{i=-m}^{+m} S_i\eta_i = 0. \tag{19}$$

Using (19), subtracting the last equation in (18) from the first two and solving for $-b$ and a' we find

$$-b = \Delta^{-1}(\Sigma\eta_i\xi_i\Sigma S_i\xi_i), \quad a' = \Delta^{-1}(\Sigma(S_i - \bar{S})^2\Sigma\eta_i\xi_i), \tag{20}$$

where

$$\Delta = \Sigma(S_i - \bar{S})^2\Sigma\xi_i^2 - (\Sigma S_i\xi_i)^2, \quad \bar{S} = n^{-1}\Sigma S_i$$

and the summation is extended from $i = -m$ to $i = m$ throughout. For practical use the formulae (20) must be further simplified; in particular, we wish to avoid the calculation of the deviates η_i. We therefore introduce sum values formed directly from the y_i

$$\left.\begin{aligned} Y_j &= \tfrac{1}{2}y_0 + \sum_{i=1}^{j-1} y_i + \tfrac{1}{2}y_j \quad \text{for} \quad j \geqslant 1, \\ Y_0 &= 0, \\ Y_j &= -\tfrac{1}{2}y_0 - \sum_{i=-1}^{j+1} y_i - \tfrac{1}{2}y_j \quad \text{for} \quad j \leqslant -1, \\ Y &= \Sigma y_i, \end{aligned}\right\} \tag{21}$$

and note the following relationships:

$$\Sigma\xi_i S_i = -\tfrac{1}{2}\Sigma\eta_i\xi_i^2, \tag{22}$$

$$\Sigma(S_i - \bar{S})^2 = \Sigma Y_i^2 + Yn^{-1}\Sigma\eta_i\xi_i^2 - Y^2n^{-2}\Sigma\xi_i^2 - n^{-1}(\Sigma Y_i)^2. \tag{23}$$

$$\Sigma S_i = \Sigma Y_i. \tag{24}$$

With the help of these relations we reach the following working formulae for the estimators a', $-b$ and c' in the order of computation. From data calculate Y_i, ΣY_i and Y (see (21)),

$$\Sigma y_i\xi_i, \quad \Sigma y_i\xi_i^2, \quad \Sigma Y_i^2, \quad \Sigma\xi_i^2 = \tfrac{1}{3}m(m+1)(2m+1).$$

Compute

$$\left.\begin{aligned} &\Sigma\eta_i\xi_i^2 = \Sigma y_i\xi_i^2 - Yn^{-1}\Sigma\xi_i^2, \\ &\Sigma(S_i - \bar{S})^2 \text{ from (23)}, \\ &\Delta = \Sigma\xi_i^2\Sigma(S - \bar{S})^2 - \tfrac{1}{4}(\Sigma\eta_i\xi_i^2)^2, \\ &a' = \Delta^{-1}\Sigma y_i\xi_i\Sigma(S - \bar{S})^2, \\ &-b = \tfrac{1}{2}\Delta^{-1}\Sigma y_i\xi_i\Sigma\eta_i\xi_i^2, \\ &c' = n^{-1}b\Sigma Y_i. \end{aligned}\right\} \tag{25}$$

Finally, we note the relation between the estimators a', $-b$ and c' and the parameters of the original exponential law, i.e. the limiting response $\hat{y}$, the exponential curvature k and f. We have

$$\hat{y} = a'/b + n^{-1}Y, \quad f\hat{y} = a'/b - c', \quad k = -2\tanh^{-1}\tfrac{1}{2}b. \tag{26}$$

As an illustration and check of these formulae we apply them in example 1 below to the theoretical sequence $y_i = 1 - e^{-x_i}$, $x_i = 0(1)4$. The table below is self explanatory, the worl following the computational order of equations (25).

Example 1. *Theoretical series* $y_i = 1 - e^{-x_i}$; *estimation of parameters by internal least square*

x_i	y_i	$2Y_i$ (from (21))	ξ_i	ξ_i^2
0	0·000	−2·129	−2	4
1	0·632	−1·497	−1	1
2	0·865	0	0	0
3	0·950	1·815	1	1
4	0·982	3·747	2	4

$Y = 3{\cdot}429$ $\qquad 1{\cdot}936 = 2\Sigma Y_i$

$$\Sigma y_i\xi_i = 2{\cdot}282, \qquad \Sigma\xi_i^2 = 10, \qquad \Sigma Y_i^2 = 6{\cdot}0270,$$
$$\Sigma y_i\xi_i^2 = 5{\cdot}510,$$
$$\Sigma\eta_i\xi_i^2 = -1{\cdot}350, \quad \Sigma(S_i - \bar{S})^2 = 0{\cdot}2119, \quad \Delta = 1{\cdot}665,$$
$$a' = 0{\cdot}290, \quad b = 0{\cdot}925, \quad c' = 0{\cdot}179,$$
$$\hat{y} = 1{\cdot}000, \quad k = -1{\cdot}000, \quad f\hat{y} = 0{\cdot}135 = e^{-2}.$$

The true values of the parameters are, of course, exactly reproduced.

(2·12) *The case* $n = even = 2m$

The only modifications of the previous section consist in defining

$$\xi_i = \tfrac{1}{2}(2i-1) \quad \text{for} \quad i = 1, \ldots, m; \qquad \xi_i = \tfrac{1}{2}(2i+1) \quad \text{for} \quad i = -1, \ldots, -m \tag{27}$$

and

$$\left.\begin{aligned} Y_j &= \sum_{i=1}^{j-1} y_i + \tfrac{1}{2}y_j \quad \text{for} \quad j \geqslant 1, \\ Y_j &= -\sum_{i=-1}^{j+1} y_i - \tfrac{1}{2}y_j \quad \text{for} \quad j \leqslant -1, \end{aligned}\right\} \tag{28}$$

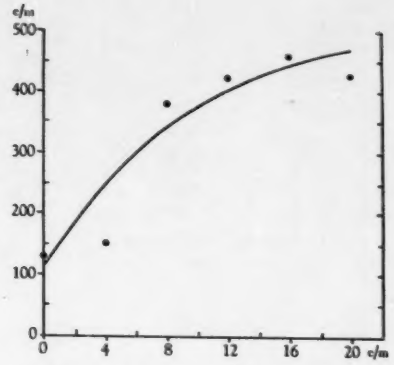
and similarly for the S_j. All other definitions, formulae and results are the same as for odd n except that the summations are from $i = -m$ to -1 and from $+1$ to $+m$, omitting 0.

Below we illustrate the procedure by fitting the exponential law of diminishing returns to yield data obtained in manurial trials of the National Agricultural Research Bureau, China.† Both trials give the response of a number of wheat varieties to the application of sulphate of ammonia at the rates of 4, 8, 12, 16 and 20 c./m.‡ The responses are grain yields and we confine ourselves here to the mean varietal responses. In the first experiment (example 2) it will be seen that there is a marked retardation of the response to the higher

† The data were kindly put at my disposal by Dr H. L. Richardson of the Imperial Chemical Industries Ltd. Many authors have applied the exponential law to the analysis of fertilizer trials (see, for example, Crowther & Yates (1941)).

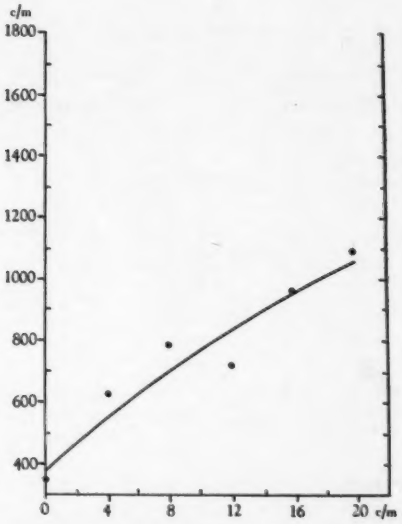
‡ c./m. = cattys per mou, the Chinese measure of yield and rate of fertilizer application.

fertilizer rates. In the second experiment (example 3), there is hardly any retardation and the estimate of the limiting response is, consequently, very high. The fit is illustrated in Figs. 1 and 2.



limiting response, $\hat{y} = 530$ c/m.

Fig. 1. Wheat yield; response to rate of fertilizer (example 2).



limiting response, $\hat{y} = 1786$ c/m.

Fig. 2. Wheat yield; response to rate of fertilizer (example 3).

Normally one would, of course, require more than six observations for fitting a three parameter regression. However, the estimation of exponential curvature and limiting response in short series is akin to the calculation of the 'quadratic effect' customary in fertilizer trials with only three levels.

Example 2. N-*fertilizer trial with wheat, Hopen, Tinghsien,* 1936. *Fit of exponential law of diminishing returns to mean varietal responses*

Fertilizer rate in c./m.	Response $y_i = 10 \times$ yield $- 1500$	$2Y_i$	$2\xi_i$	$4\xi_i^2$
0	127	−1187	−5	25
4	151	− 909	−3	9
8	379	− 379	−1	1
12	421	421	1	1
16	460	1302	3	9
20	426	2188	5	25

$Y = 1964$ 1436 Check:† $1436 = 6(1307 - 657) - 2(1232)$
$Y^- = 657$*
$Y^+ = 1307$

$\Sigma y_i \xi_i = 1232$, $\Sigma \xi_i^2 = 17{\cdot}5$, $\Sigma Y_i^2 = 225\,9670$,
$\Sigma y_i \xi_i^2 = 5031$, $\Sigma Y_i = 718$,
$\Sigma \eta_i \xi_i^2 = -697$, $\Sigma (S_i - \bar{S})^2 = 70524$, $\Delta = 111\,2718$,
$a' = 78{\cdot}1$, $b = +0{\cdot}386$, $c' = 462$,
$\hat{y} = 530$, $k = -0{\cdot}391$, $\hat{y}f = 156$.

Fitted law: $y_i = 529{\cdot}6 - 156{\cdot}1 \exp(-0{\cdot}391 \xi_i)$.

Example 3. N-*fertilizer trial with wheat, Shantung,* 1936. *Fit of exponential law of diminishing returns to mean varietal responses*

Fertilizer rate in c./m.	Response $y_i = 10$ yield $- 3000$	$2Y_i$	$2\xi_i$	$4\xi_i^2$
0	353	−3179	−5	25
4	627	−2199	−3	9
8	786	− 786	−1	1
12	717	717	1	1
16	959	2393	3	9
20	1093	4445	5	25

$Y = 4535$ 1391 Check: $1391 = 6(2769 - 1766) - 2(2313{\cdot}5)$
$Y^- = 1766$
$Y^+ = 2769$

$\Sigma y_i \xi_i = 2313{\cdot}5$, $\Sigma \xi_i^2 = 17{\cdot}5$, $\Sigma Y_i^2 = 103\,89\,500$,
$\Sigma y_i \xi_i^2 = 12982$, $\Sigma Y_i = 695{\cdot}5$,
$\Sigma \eta_i \xi_i^2 = -245$, $\Sigma (S_i - \bar{S})^2 = 126230$, $\Delta = 219\,4019$,
$a' = 133$, $b = 0{\cdot}1292$, $c' = 15{\cdot}0$,
$\hat{y} = 1786$, $k = -0{\cdot}1294$, $\hat{y}f = 1015$.

Fitted law: $y_i = 1786 - 1015 \exp(-0{\cdot}1294 \xi_i)$.

† It is convenient, when adding the y_i, to record $Y^- = \Sigma y_i$ for $i < 0$ and $Y^+ = \Sigma y_i$ for $i > 0$, separately (forming $Y = Y^- + Y^+ = \Sigma y_i$). This provides a check on the forming and copying of the Y_i from the equation $\Sigma 2Y_i = n(Y^+ - Y^-) - 2\Sigma y_i \xi_i$, e.g. in the present case $6(1307 - 657) - 2(1232) = 1436$.

We now turn to two well-known regression laws which, by suitable transformation, can be reduced to our exponential law.

(2·2) *The internal least square fit of the logistic curve*

The general form of the logistic curve is†

$$z = A/(1+Be^{-kx}). \tag{29}$$

In the case where A is known (as for instance with a biological response known to vary between 0 and 100 %) the transformation

$$\log(A/z-1) = \log B - kx, \tag{30}$$

reduces the problem to a plain linear regression fit for $\log B$ and k. Nevertheless, this case has recently been dealt with (Finney, 1947) by the more elaborate, but under certain conditions more efficient, method of maximum likelihood.

We are here dealing with the general case where A has to be estimated from the data. In this case the transformation $z = 1/y$ yields

$$y = 1/A + (B/A)e^{-kx}, \tag{31}$$

which is of the form (1) or (9). Considerations of appropriateness and efficiency are, again, postponed for discussion in § 3 below, but we should mention here that Rhodes (1940) also uses reciprocals when fitting the logistic by a different method. We are giving in example 4, below, the data used by him, viz. Population figures for the U.S. Census Data, 1800–1910. It is with such data on population growth that the estimation of the logistic parameters is of importance.

Example 4. *U.S. Census population* 1800–1910. *Fit of logistic curve and estimation of its parameters by internal least squares*

Year x	Population (millions) z	$y = 10{,}000/z$	$2Y_i$	$2\xi_i$	$4\xi_i^2$
1800	5·308	1884	−10310	−11	121
10	7·240	1381	− 7045	− 9	81
20	9·638	1038	− 4626	− 7	49
30	12·866	777	− 2811	− 5	25
40	17·069	586	− 1448	− 3	9
1850	23·192	431	− 431	− 1	1
60	31·443	318	318	1	1
70	38·558	259	895	3	9
80	50·156	199	1353	5	25
90	62·948	159	1711	7	49
1900	75·995	132	2002	9	81
10	91·972	109	2243	11	121

$Y = 7273$
$Y^- = 6097$
$Y^+ = 1176$

† In some applications it has been found necessary to generalize the logistic to

$$y - C = A/(1+B\exp(-kx)),$$

where the lower asymptote C is a fourth parameter to be estimated from the sample. However, this case is comparatively rare.

Following the formulae and procedure (25), we obtain

$$a' = -172{\cdot}15, \quad b = 0{\cdot}3074, \quad c' = -232{\cdot}46,$$

whence

$$\hat{y} = 50{\cdot}4, \quad k = -0{\cdot}3098, \quad f\hat{y} = -323{\cdot}4.$$

By comparison with (31) we obtain from the above parameter estimates of the exponential law, the corresponding estimates for the logistic. These are given below, alongside those obtained by Rhodes, who used a method which is computationally more cumbersome as it requires both the reciprocal, as well as a logarithmic transformation.

Example 4. *U.S. Census data. Comparison of estimates of logistic parameters obtained by Rhodes with those obtained by internal regression*

Parameter	Internal regression	Rhodes' (1940) method
$10{,}000\hat{f\hat{y}}$	198	199
$-k$	0·3098	0·3127
$-\hat{\hat{y}}f$	323	328

The agreement is extremely close, which is to be expected as the logistic curve fits the data well. In general, the theoretical properties of Rhodes's method are difficult to assess as it consists of two fitting procedures. The first (which is akin to a lag-1 autoregressive scheme) provides the estimates of k and A; these are then used to transform to linearity (as in (30)), so that the estimate of B/A is obtained as a mean of the log transforms.

(2·3) *The internal least square fit of the Makeham-Gompertz curve*

This curve is of fundamental importance in Life Table work. Its general form is

$$z = a\exp(-be^{-kx}). \tag{32}$$

Using the transformation $y = \log z$, we obtain

$$y = \log a - be^{-kx}, \tag{33}$$

which is of the form (1) or (9), so that its parameters can be estimated by the same method. We intend to deal with this application more fully elsewhere.

(2·4) *The special case of the exponential law with known asymptote (Markoff chain)*

A particular case of the first order linear difference equation is the case $a = 0$. In this case, (8) is equivalent to

$$y_{i+1} - y_i = -\tfrac{1}{2}b(y_{i+1} + y_i), \tag{34}$$

which is the systematic part of the Markoff chain equation

$$y_{i+1} = \mu y_i + \nu_i, \tag{35}$$

in which ν_i represents a random disturbance. In analogy to the general case, we have as the solution of (34)

$$y_i = y^* e^{kxi}, \tag{36}$$

where

$$k = -2\tanh\tfrac{1}{2}b, \quad y^* = \hat{y}f. \tag{37}$$

Dealing with the case $n = \text{odd} = 2m+1$, retaining the definitions (21), we find by summation of (34)

$$\eta_j = -bY_j + w, \tag{38}$$

where w is a constant of integration. Obviously (34) and (38) are, again, equivalent equations. The least square solutions for b and w in (38) are then given by

$$\Sigma\eta_i Y_i = -b\Sigma Y_i^2 + w\Sigma Y_i, \quad \Sigma\eta_i = -b\Sigma Y_i + wn = 0. \tag{39}$$

From the second equation it follows that $w = bn^{-1}\Sigma Y_i$ and hence, from the first equation, that

$$-b = \Sigma Y_i\eta_i / \Sigma(Y_i - \overline{Y})^2. \tag{40}$$

By partial summation we may transform (40) into

$$-b = n^{-1}\Sigma\eta_i\xi_i\Sigma y_i / \Sigma(Y_i - \overline{Y})^2,$$

which is our least square estimate of b. It is proportional to the mean of the y and to the regression coefficient y on x, whilst it is inversely proportional to the sum of squares of the progressive totals Y_i of the y_i. The computational procedure is simplified as $\Sigma\eta_i\xi_i^2$ is no longer required.

3. LARGE SAMPLE THEORY OF THE INTERNAL LEAST SQUARE ESTIMATORS

(3·1) *The relation between internal least square and ordinary least square estimates*

In this section we compare the basic assumption on which the internal least square method is based with that made in the ordinary least square hypothesis. We confine ourselves to the special case of an exponential law with x axis as asymptote, as in (36). The general case follows on the same lines. If we denote the residuals of (36) by ϵ_i, viz.

$$\epsilon_i = y_i - y^* e^{kx_i}, \tag{41}$$

then the assumption on which the classical least square fit is based, is that the ϵ_i are independent and have the same variance. The roots y^* and k of, $\Sigma\epsilon_i^2 =$ minimum, are then maximum likelihood estimates. If on the other hand we assume that the residuals of (39), viz.

$$\zeta_i = \eta_i + b(Y_i - \overline{Y}), \tag{42}$$

are independent and have the same variance, then the internal least square estimators have the maximum likelihood property. Yet another hypothesis can be investigated:

Analogous to Bartlett's (1946) treatment of stationary time series in correlogram work, we may consider the residuals

$$\theta_i = (y_{i+1} - y_i) + \tfrac{1}{2}b(y_{i+1} + y_i) \tag{43}$$

in the difference equation (34). The assumption of independent deviates, θ_i, which is the analogue to Bartlett's starting point, would lead in the present case to the estimator

$$-\tfrac{1}{2}b = (y_m^2 - y_{-m}^2)/\Sigma(y_{i+1} + y_i)^2. \tag{44}$$

This is unsuitable, as its numerator depends on the first and last observation only. It is interesting to note the relation between the ϵ_i and ζ_i of (41) and (42). We find:

$$\zeta_i = \epsilon_i + b(E_i - \overline{E}), \tag{45}$$

where E_i and $\overline{E}$ are formed from the ϵ_i on the lines of (21). Since b is of the order n^{-1} and therefore small,† it follows that the second term in (45) is small compared with the first, particularly near the centre of the range $j = 0$. Thus the assumption of independent ζ_i does not differ seriously from that of the classical least square method. It is unlikely, therefore, that many situations will arise in which the assumption of independent ζ_i can be disproved

† It must be remembered that Y_m is a total of $\frac{1}{2}n$ values of y_i and hence, from (38), $b \sim 2\overline{y}/n\overline{y}$.

whilst the hypothesis of independent ϵ_i can be accepted. Nevertheless, we give in the next section the loss in efficiency resulting from the use of the internal least-square-fit under the more usual hypothesis that the ϵ_i are independent deviates with equal variance.

(3·2) *The efficiency of the internal least square estimator*

In this section we confine ourselves to the case $a = 0$, i.e. the exponential law with x axis as asymptote. The treatment of the general first order internal regression is on similar lines but algebraically more tedious. Further, in order to simplify the formulae we approximate exponential sums of the type $h \sum_{i=-m}^{+m} e^{kx_i} f(x_i)$ with $x_{i+1} - x_i = h$ and $n = 2m+1$, by the integrals

$$\int_{-\frac{1}{2}X}^{+\frac{1}{2}X} e^{kx} f(x)\, dx,$$

with $X = nh$. The error thereby committed is small, provided n is moderate or large. The exact summations could be carried out, but are more tedious.

We first derive the transcendental equations for the ordinary least square estimators and derive their variances from the general maximum likelihood formulae. We then derive an approximate large sample formula for the variance of the internal least square estimator b, and finally by comparison with the former, its efficiency.

(3·21) *Maximum likelihood results*

Let the range of the independent variate be $-\frac{1}{2}X \leqslant x \leqslant \frac{1}{2}X$; $X = nh$, and let us try to fit the expression

$$y(x) = y^* e^{kx} \tag{46}$$

to the observed series $y(x_i)$ where $x_i = ih$, $h = X/n$. Then, by ordinary least square or maximum likelihood procedure, we minimize

$$L = X^{-1} \int_{-\frac{1}{2}X}^{+\frac{1}{2}X} (y - y^* e^{kx})^2\, dx, \tag{47}$$

resulting in the nonlinear system of equations

$$\left.\begin{aligned} \frac{\partial L}{\partial k} &= X^{-1} \int_{-\frac{1}{2}X}^{+\frac{1}{2}X} (y - y^* e^{kx})\, e^{kx} \times y^*\, dx = 0, \\ \frac{\partial L}{\partial y^*} &= X^{-1} \int_{-\frac{1}{2}X}^{+\frac{1}{2}X} (y - y^* e^{kx})\, e^{kx}\, dx = 0. \end{aligned}\right\} \tag{48}$$

In order to obtain the variance of the maximum likelihood estimator of k ($\hat{k}$ say) we must form the Hessian

$$\Delta = \begin{vmatrix} E\left(\frac{\partial^2 L}{\partial k^2}\right) & E\left(\frac{\partial^2 L}{\partial k\, \partial y^*}\right) \\ E\left(\frac{\partial^2 L}{\partial k\, \partial y^*}\right) & E\left(\frac{\partial^2 L}{\partial y^{*2}}\right) \end{vmatrix} \tag{49}$$

We obtain

$$\Delta = \begin{vmatrix} \frac{1}{8} X^{-1} y^2 k^{-3} \{(2+q^2)(e^q - e^{-q}) - 2q(e^q + e^{-q})\}, & \frac{1}{4} X^{-1} y k^{-2} \{q(e^q + e^{-q}) - (e^q - e^{-q})\} \\ \frac{1}{4} X^{-1} y k^{-2} \{q(e^q + e^{-q}) - (e^q - e^{-q})\}, & \frac{1}{2} X^{-1} k^{-1} (e^q - e^{-q}) \end{vmatrix}, \tag{50}$$

where $q = kX$ and k is the true parameter. Accordingly we find for the variance of $\hat{k}$ (which is given by the ratio of first minor to n times the determinant)

$$\text{Variance } \hat{k} = \sigma_\epsilon^2 8 y^{*-2} q^3 n^{-1} X^{-2} (e^q - e^{-q}) / (4q^2 + (e^q - e^{-q})^2), \tag{51}$$

which can be reduced to

$$\text{Variance } \hat{k} = \sigma_\epsilon^2 8y^{*-2}q^3 n^{-1}X^{-2}\sinh q/(\cosh 2q - 2q^2 - 1). \tag{52}$$

Letting $k \to 0$, but keeping X fixed (i.e. letting $q \to 0$) we have

$$\text{Variance } \hat{k} \to \sigma_\epsilon^2 12y^{*-2}n^{-1}X^{-2}. \tag{53}$$

As an independent check on (53) we remember that for $q \to 0$ with y^* large and fixed, the exponential law (46) will be approximately represented by the line

$$y(x) = y^* + y^* kx. \tag{54}$$

The maximum likelihood estimator of y^*k, that is the ordinary linear regression coefficient, has a variance $\sigma_\epsilon^2/\Sigma(x_i - \bar{x})^2$ which, for the present sample of x_i equidistantly spread over the interval of length X, tends to $\sigma_\epsilon^2 12/(X^2 n)$, thereby confirming (53).

(3·22) *Internal least square results*

We now turn to the internal least square estimator b and study its random sampling distribution under the hypothesis that the observed values of y_i have expectations $y^* \exp(kx_i)$, from which they differ by independent random deviates ϵ_i having equal variances viz.

$$y(x_i) = y^* \exp(kx_i) + \epsilon_i. \tag{55}$$

We again use the approximation $Xn^{-1}\Sigma f(x_i) \cong \int_{-\frac{1}{2}X}^{+\frac{1}{2}X} f(x)\,dx$.

Using elementary and partial integration and putting $kX = q$, we reach the following results:

$$\left.\begin{aligned} \text{Total of } y = \int_{-\frac{1}{2}X}^{+\frac{1}{2}X} y(x)\,dx &= \frac{y^*}{k}(e^{\frac{1}{2}q} - e^{-\frac{1}{2}q}) + \int_{-\frac{1}{2}X}^{+\frac{1}{2}X} \epsilon\,dx, \\ \bar{y} = X^{-1}\,\text{total of } y &= y^* q^{-1}(e^{\frac{1}{2}q} - e^{-\frac{1}{2}q}) + \bar{\epsilon}; \end{aligned}\right\} \tag{56}$$

where

$$\left.\begin{aligned} \eta &= y - \bar{y} = y^* e^{kx} + k\gamma + \epsilon - \bar{\epsilon}, \\ \gamma &= -y^*(qk)^{-1}(e^{\frac{1}{2}q} - e^{-\frac{1}{2}q}); \end{aligned}\right\} \tag{57}$$

$$Y(x) = \int_0^x y(x)\,dx = y^* k^{-1} e^{kx} + \int_0^x \epsilon\,dx; \tag{58}$$

$$Y - \overline{Y} = y^* k^{-1} e^{kx} + \gamma + \int_0^x \epsilon\,dx - X^{-1}\int_{-\frac{1}{2}X}^{+\frac{1}{2}X}\int_0^x \epsilon\,d\xi\,dx; \tag{59}$$

$$\int_{-\frac{1}{2}X}^{+\frac{1}{2}X} (Y - \overline{Y})^2\,dx = A + 2B\int_{-\frac{1}{2}X}^{+\frac{1}{2}X} \epsilon\,dx - 2\int_{-\frac{1}{2}X}^{+\frac{1}{2}X} \epsilon C(x)\,dx; \tag{60}$$

where

$$\left.\begin{aligned} A &= \tfrac{1}{2}y^{*2}k^{-3}(e^q - e^{-q}) - \gamma^2 X, \\ B &= \tfrac{1}{2}y^* k^{-2}(e^{\frac{1}{2}q} + e^{-\frac{1}{2}q}), \end{aligned}\right\} \tag{61}$$

and

$$C(x) = y^* k^{-2} e^{kx} + \gamma x,$$

and where the quadratic term in ϵ has been ignored. Similarly we obtain:

$$\int_{-\frac{1}{2}X}^{+\frac{1}{2}X} \eta Y\,dx = k\left(A + B\int_{-\frac{1}{2}X}^{+\frac{1}{2}X} \epsilon\,dx + \int_{-\frac{1}{2}X}^{+\frac{1}{2}X} \epsilon(-C(x) + D(x))\,dx\right), \tag{62}$$

where

$$D(x) = y^* k^{-2} e^{kx} + k^{-1}\gamma,$$

and where, again, the quadratic term in ϵ has been ignored. We now form our estimator b in accordance with (40):

$$-b = \int_{-\frac{1}{2}X}^{+\frac{1}{2}X} \eta Y\,dx \Big/ \int_{-\frac{1}{2}X}^{+\frac{1}{2}X} (Y - \overline{Y})^2\,dx, \tag{63}$$

and obtain to within linear terms in ϵ:

$$-b = k\left(1 + A^{-1}\int_{-\frac{1}{2}X}^{+\frac{1}{2}X} \epsilon(-B + C(x) + D(x))\,dx\right). \tag{64}$$

We first note that to within the approximation employed the expectation of $-b$ is k, which is the first term in the expansion of $-b = 2\tanh\frac{1}{2}k$.†

Next we see that we have a representation of $-b$ in the form constant $+\int \epsilon f(x)\,dx$, i.e. a weighted sum of residuals. Remembering now that $\int \epsilon f\,dx$ is being used as an approximation to $\frac{X}{n}\Sigma f(x_i)\epsilon_i$ and using, therefore, the formula

$$\text{Variance}\left(\int_{-\frac{1}{2}X}^{+\frac{1}{2}X} \epsilon f(x)\,dx\right) = \sigma_\epsilon^2 X n^{-1}\int_{-\frac{1}{2}X}^{+\frac{1}{2}X} f^2(x)\,dx, \tag{65}$$

we reach

$$\text{Variance}\, b = \sigma_\epsilon^2 X n^{-1}\int_{-\frac{1}{2}X}^{+\frac{1}{2}X} (-B + C(x) + D(x))^2\,dx, \tag{66}$$

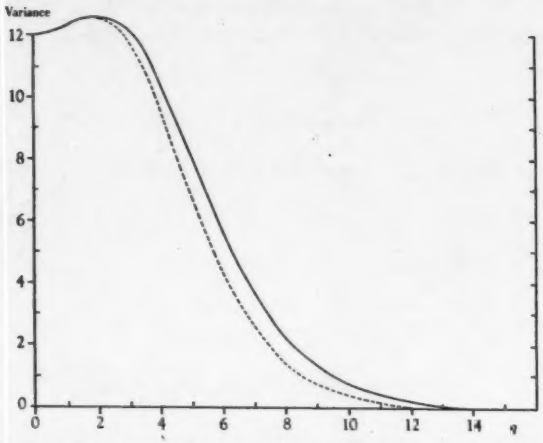


Fig. 3. Comparison of the variance § of alternative estimators:
(*a*) Maximum likelihood (equation 52) --------
(*b*) Internal least square (equation 67) ————

§ The variances are standardized by putting $\sigma_\epsilon^2 = 1$, $X = 1$, $n = 1$, $y^* = 1$

which, after some lengthy algebra, emerges as

$$\text{Variance}\,(b) = \sigma_\epsilon^2 \frac{4q^3}{nX^2y^{*2}} \frac{-1 + \left(\frac{1}{q} + \frac{q}{12}\right)\tanh\frac{1}{2}q + \frac{1}{4}q\coth\frac{1}{2}q}{\left(1 - \frac{2}{q}\tanh\frac{1}{2}q\right)^2 (e^q - e^{-q})}. \tag{67}$$

It is easy to see that for $q \to 0$, $\text{var}\, b \to 12\sigma_\epsilon^2/nX^2y^{*2}$, which is the same limit as (53) and therefore agrees with both the maximum likelihood value as well as the classical variance of the linear regression coefficient.

† For the order of discrepancy between $-b$ and k see, for instance, examples 1 to 4 of Part 2.

(3·23) *The efficiency of* b

In Fig. 3 we have plotted both the maximum likelihood variance of $\hat{k}$ as well as that of b. For given values of σ_ϵ^2, n, X and y^*, both variances are functions of $q = kX$. We have already seen that for $q \to 0$ they tend to the same limit, so that b is highly efficient for small q,† which was to be expected from the discussion given in § (3·1). As q increases both variances decrease rapidly, but the maximum likelihood variance more so and the efficiency of b drops until for large q the efficiency tends to $600/q$ %. For such large values of q (say $q \geqslant 10$), when the fitted exponential is steep and sharply bent and when $y_{-m} \geqslant 10y_m$, it is doubtful whether the basic assumption of uniform variance of the residuals ϵ_i over the whole range X is justified.

The objection that the variance of b depends on q, i.e. the parameter estimated by $-bX$, can of course be overcome (at least approximately) by an appropriate variate transformation of b. We shall not enter into this problem here.

REFERENCES

BARTLETT, M. S. (1946). *J. Roy. Statist. Soc. Suppl.* **8**, 27.
CROWTHER, E. M. & YATES, F. (1941). *Empire J. Exp. Agric.* **9**, 77.
CUNNINGHAM, L. B. C. & HYND, W. R. B. (1946). *J. Roy. Statist. Soc: Suppl.* **8**, 62.
FINNEY, D. J. (1947). *J. Roy. Statist. Soc. Suppl.* **9**, 46.
KENDALL, M. G. (1944). *Biometrika*, **33**, 105.
MANN, H. B. & WALD, A. (1943). *Econometrica*, **11**, 173.
RHODES, E. C. (1940). *J. Roy. Statist. Soc.* **103**, 362.
SASULY, M. (1934). *Trend Analysis of Statistics*. Brookings Institution, Washington, D.C.

† Actually, on substituting in equations (52) and (67) it will be found that near $q = 2$ the maximum likelihood variance is very slightly in excess of that of the internal least square solution, an anomaly caused by the approximations on which (67) is based.

THE GEOMETRICAL METHOD IN THE THEORY OF SAMPLING

By DAVID FOG, *den kgl. Veterinaer- og Landbohøjskole, Copenhagen*

For the determination of exact sampling distributions a number of different methods have been employed, of which two will be mentioned here:

(1) The geometrical method, which has been applied with great success especially by R. A. Fisher (1915, 1925) and consists in the use of the terminology and results of multi-dimensional geometry;

(2) the analytical method, the main line of which is a change of variables and calculation of the corresponding functional determinants.

The first method appears elegant and short, whereas the second frequently leads to long calculations. On the other hand, the view is often expressed that the geometrical method arrives at its results too easily, at the cost of the accuracy usually required, and which is duly honoured by the analytical method. Probably this view has been affirmed by the circumstance that it has proved difficult to translate the geometrical methods and considerations so as to enter naturally and smoothly into the usual analytical treatment.

With regard to the criticism that the geometrical method is less accurate, it should be observed that the justification depends entirely on the way in which the multidimensional geometry is built up. The construction of the multidimensional geometry may be worked out on a purely numerical basis, without introducing a single geometrical axiom, but in such a manner that the meaning of the geometrical terms as well as the contents of the geometrical formulae are of an entirely numerical nature. The system then rests on the same basis as the usual mathematical analysis and is therefore just as precise as the latter.

This subject will not be discussed further here, but we shall show in the present paper how in a number of cases the geometrical method may, with complete preservation of its simplicity, be translated into analytical form, so that, without any knowledge of multi-dimensional geometry, advantage may be taken of the simple methods to which it has given rise.

Before dealing with the examples we shall recall the concepts of volume and surface-content of an n-dimensional sphere. Such a sphere lies in an n-dimensional space and may be represented by the equation

$$x_1^2+x_2^2+\ldots+x_n^2=a^2; \tag{1}$$

its interior is the domain

$$x_1^2+x_2^2+\ldots+x_n^2<a^2. \tag{2}$$

The volume of this domain is determined as the multiple integral

$$V_n(a)=\int\limits_{\sum_1^n x_i^2<a^2}\ldots\ldots\int dx_1 dx_2\ldots dx_n, \tag{3}$$

and it is easily shown that

$$V_n(a)=\frac{\pi^{\frac{1}{2}n}a^n}{\Gamma(\frac{1}{2}n+1)}. \tag{4}$$

In order to find the corresponding surface-content $S_n(a)$ we consider $dV_n(a)$. This denotes the volume between two concentric spheres of radii a and $a+da$; thus $dV_n(a) = S_n(a)\,da$. We therefore obtain

$$S_n(a) = \frac{dV_n(a)}{da} = \frac{2\pi^{\frac{1}{2}n}a^{n-1}}{\Gamma(\frac{1}{2}n)}. \tag{5}$$

It should be noted that the result (4), even with exclusion of all geometry, has a definite meaning as the value of the multiple integral (3). As for (5), from a purely analytical point of view, we need only regard it in the following as a definition of $S_n(a)$.

We shall now consider three examples. The first two are quite elementary, and the results of all three are well known. Each example will be treated in two ways, in a geometrical and in an analytical manner, the latter being the translation of the former into analytical language; the essential points are the possibility of this translation and its method of working out.

Example 1. Let $x_1, x_2, \ldots, x_n$ be n mutually independent variables, all normal $(0, \sigma)$, i.e. having the frequency function

$$p\{x\} = \frac{1}{\sqrt{(2\pi)}\sigma}\exp\left[-\frac{x^2}{2\sigma^2}\right]. \tag{6}$$

The frequency function of the set $X = (x_1, x_2, \ldots, x_n)$ is then

$$\kappa = p\{X\} = \frac{1}{(\sqrt{(2\pi)}\sigma)^n}\exp\left[-\frac{\sum_1^n x_i^2}{2\sigma^2}\right]. \tag{7}$$

Putting

$$q = \sqrt{\left(\sum_1^n x_i^2\right)}, \tag{8}$$

the formula (7) may also be written

$$\kappa = \frac{1}{(\sqrt{(2\pi)}\sigma)^n}\exp\left[-\frac{q^2}{2\sigma^2}\right]. \tag{9}$$

Assuming X to be a point in an n-dimensional space and $\kappa = p\{X\}$ to be the density at this point, the probability of X being in a domain ω is equal to the mass of ω and is given by the multiple integral

$$\int\ldots\int \kappa\, dX, \tag{10}$$

taken throughout ω, where $dX = dx_1 dx_2 \ldots dx_n$ is an element of volume.

This in particular comprises the determination of the distribution of an arbitrary statistic

$$y = \phi(x_1, x_2, \ldots, x_n).$$

Designating its distribution function as $P\{y\}$, (10) gives $dP\{y\}$, when for ω we put the domain

$$y < \phi(x_1, x_2, \ldots, x_n) < y + dy.$$

We shall now find the distribution of the statistic q, introduced in (8). This variable may be interpreted as the distance from the origin O to the point X. Hence we use as element of volume the region between two n-dimensional spheres of common centre O and radii q and $q+dq$ and obtain

$$dP\{q\} = \kappa\, dV_n(q) = \kappa S_n(q)\, dq.$$

Inserting the values of κ and $S_n(q)$ from (9) and (5) we get

$$dP\{q\} = \frac{1}{2^{\frac{1}{2}(n-2)}\Gamma(\frac{1}{2}n)}\exp\left[-\frac{q^2}{2\sigma^2}\right]\left(\frac{q}{\sigma}\right)^{n-1}\frac{dq}{\sigma}, \tag{11}$$

and we have obtained the distribution required. For $\sigma = 1$ (11) in particular shows the distribution of χ.

Next we will prove (11) without any use of geometry, but following the exposition above as closely as possible. We then start with (9)—where q is determined by (8)—and

$$dP\{X\} = \kappa dX. \tag{12}$$

By an almost* one-to-one correspondence we pass from the variables $X = (x_1, x_2, \ldots, x_n)$ to

$$q, \quad U = (u_1, u_2, \ldots, u_{n-1}),$$

the u's being co-ordinates on the unit-sphere, the specification of which is of no importance for the proof.† Designating by Δ the jacobian $\partial(X)/\partial(q, U)$, we get

$$dP\{q, U\} = \kappa \,|\, \Delta \,|\, dq dU,$$

from which we find

$$dP\{q\} = \kappa dq \int\ldots\int_{\omega} |\, \Delta \,|\, dU, \tag{13}$$

ω being the domain of the u's.

Hence it remains only to determine the multiple integral in (13), which is done as follows:

From

$$V_n(a) = \int\ldots\ldots\int_{\sum_1^n x_i^2 < a^2} dX$$

we obtain by means of the substitution used above

$$V_n(a) = \int\ldots\int |\, \Delta \,|\, dq dU = \int_0^a dq \int\ldots\int_{\omega} |\, \Delta \,|\, dU,$$

thus by differentiation

$$\frac{dV_n(q)}{dq} = \int\ldots\int_{\omega} |\, \Delta \,|\, dU.$$

Hence the integral desired is equal to $S_n(q)$, and (13) becomes

$$dP\{q\} = \kappa S_n(q)\, dq,$$

which, just as before, leads to (11).

Note. The assumption that the true mean of x is 0 is of little importance. For if this mean is equal to $\mu \neq 0$, we have only to consider, instead of x, the deviation $x - \mu$ from μ. We then find that

$$q = \sqrt{\left(\sum_1^n (x_i - \mu)^2\right)}$$

has the distribution (11). A similar modification will apply to the following examples.

Example 2. As before, let $X = (x_1, x_2, \ldots, x_n)$ be distributed so that (6) and (7) are valid. From the x's we form the linear combination

$$z = \sum_1^n a_i x_i, \tag{14}$$

where for the sake of convenience we may assume that

$$\sum_1^n a_i^2 = 1. \tag{15}$$

* The exceptional points forming a domain of volume 0.

† A suitable transformation is for instance

$$\begin{aligned}
x_1 &= q \cos u_1 && (q > 0),\\
x_2 &= q \sin u_1 \cos u_2 && (0 < u_i < \pi,\ i = 1, 2, \ldots, n-2),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\\
x_{n-1} &= q \sin u_1 \sin u_2 \ldots \sin u_{n-2} \cos u_{n-1} && (0 < u_{n-1} < 2\pi),\\
x_n &= q \sin u_1 \sin u_2 \ldots \sin u_{n-2} \sin u_{n-1}.
\end{aligned}$$

In Fig. 1, α represents the hyperplane $\sum_1^n a_i x_i = 0$, passing through O; the line l is perpendicular to α at O. The point X is projected on l and α as N and Q; the quadrilateral $ONXQ$ is a rectangle. The co-ordinates of N and Q are easily found to be

$$N = (a_1 z, \ldots, a_n z), \quad Q = (x_1 - a_1 z, \ldots, x_n - a_n z).$$

Further $ON = z$, and for OQ, denoted by q_1, we find

$$q_1^2 = \sum_1^n (x_i - a_i z)^2. \tag{16}$$

From the right-angled triangle OQX, where $OX^2 = \sum_1^n x_i^2$, we have

$$\sum_1^n x_i^2 = z^2 + q_1^2, \tag{17}$$

which may also be easily verified directly by calculation.

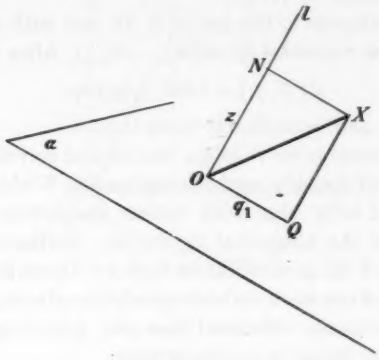


Fig. 1.

By means of (17), we may write (7) as

$$\kappa = \frac{1}{(\sqrt{(2\pi)}\sigma)^n} \exp\left[-\frac{z^2 + q_1^2}{2\sigma^2}\right]. \tag{18}$$

We will now find the joint distribution of z and q_1. As element of volume for X we use a domain, the projection of which on α is the region between two $(n-1)$-dimensional spheres in α, having the common centre O and the radii q_1 and $q_1 + dq_1$, and the projection of which on l is the line-element dz. We then get

$$dP\{z, q_1\} = \kappa dV_{n-1}(q_1)\, dz = \kappa S_{n-1}(q_1)\, dq_1\, dz. \tag{19}$$

Hence by (18) and (5)

$$dP\{z, q_1\} = \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left[-\frac{z^2}{2\sigma^2}\right] dz \frac{1}{2^{\frac{1}{2}(n-3)}\Gamma\frac{1}{2}(n-1)} \exp\left[-\frac{q_1^2}{2\sigma^2}\right] \left(\frac{q_1}{\sigma}\right)^{n-2} \frac{dq_1}{\sigma}. \tag{20}$$

Thus z and q_1 are independent; z is normal $(0, \sigma)$, and q_1 is distributed as q in (11), but with $n-1$ instead of n.

If in particular $a_1 = a_2 = \ldots = a_n = 1/\sqrt{n}$, we get, $\bar{x}$ being the mean $\sum_1^n x_i/n$,

$$z = \bar{x}\sqrt{n}, \quad q_1^2 = \sum_1^n (x_i - \bar{x})^2.$$

Let us now, without the aid of geometry, once more prove (20), starting from (6), (7), (14), (15) and

$$dP\{X\} = \kappa dX. \tag{21}$$

We introduce new variables z, $Y = (y_1, y_2, \ldots, y_{n-1})$ by means of an orthogonal substitution

$$z = a_1x_1 + a_2x_2 + \ldots + a_nx_n, \quad y_i = a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n \quad (i = 1, 2, \ldots, n-1), \tag{22}$$

and put

$$q_1 = \sqrt{\left(\sum_1^{n-1} y_i^2\right)}. \tag{23}$$

As (22) leaves sums of squares invariant we get

$$\sum_1^n x_i^2 = z^2 + q_1^2. \tag{24}$$

A comparison with (17) shows that the q_1's introduced by (23) and (16) are identical.

The jacobian corresponding to (22) being ± 1 we get from (21)

$$dP\{z, Y\} = \kappa dz dY, \tag{25}$$

where as before κ is determined by (18).

By a transformation analogous to the one on p. 48, but with $n-1$ instead of n, we may instead of Y introduce new variables $(q_1, u_1, u_2, \ldots, u_{n-2})$. After integration relative to the u's we have

$$dP\{z, q_1\} = \kappa dz S_{n-1}(q_1) dq_1,$$

which is equivalent to (19) and immediately leads to (20).

Example 3. As a final example we consider the normal correlation between k variables. In the geometrical treatment we use a method employed by Wishart (1928) for the derivation of the distribution named after him, with certain simplifications, however, which will facilitate the transition to the analytical exposition. Furthermore, for convenience, we assume in the following $k = 3$; the generalization from $k = 3$ to an arbitrary value should cause no difficulties, the addition of one more variable merely introducing another step of reduction.

Let x, y, z denote three normally correlated variables. Assuming that the three true means are equal to 0, the frequency function may be written

$$p\{x, y, z\} = \frac{\sqrt{A}}{(2\pi)^{\frac{3}{2}}} \exp\left[-\tfrac{1}{2}(a_{11}x^2 + 2a_{12}xy + \ldots + a_{33}z^2)\right], \tag{26}$$

where A is the determinant $|a_{\mu\nu}|$. We consider n triples (x_i, y_i, z_i), mutually independent and each satisfying (26). We introduce the designations

$$l_{11} = \sum_1^n x_i^2, \quad l_{12} = \sum_1^n x_i y_i, \quad \ldots, \quad l_{33} = \sum_1^n z_i^2. \tag{27}$$

Putting furthermore in the same way as before

$$X = (x_1, x_2, \ldots, x_n), \quad Y = (y_1, y_2, \ldots, y_n), \quad Z = (z_1, z_2, \ldots, z_n), \tag{28}$$

we get

$$\kappa = p\{X, Y, Z\} = \frac{A^{\frac{1}{2}n}}{(2\pi)^{\frac{3}{2}n}} \exp\left[-\tfrac{1}{2}(a_{11}l_{11} + 2a_{12}l_{12} + \ldots + a_{33}l_{33})\right]. \tag{29}$$

In the following our aim is to find the distribution of the quantities $l_{11}, l_{12}, \ldots, l_{33}$. κ in (29) denotes the density at the point

$$M = (x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n, z_1, z_2, \ldots, z_n) \tag{30}$$

in a $3n$-dimensional space; the origin is denoted O. Within this space we consider three n-dimensional subspaces, corresponding to the first n, the intermediate n, and the last n co-ordinates. The projections of M on these three subspaces are the points X, Y, and Z in (28)

in the sense that for each point only the n co-ordinates belonging to the corresponding subspace are indicated (the other $2n$ being zero).

For the time being it will be convenient to imagine that Y and Z, with the co-ordinates shown in (28), are placed in the same n-dimensional subspace as X. Thereby we are, *inter alia*, enabled to give a simple geometrical interpretation of all the l's. Thus l_{11} means the square of the distance OX, while l_{12} means the product $OX.OY.\cos\phi_{12}$, where ϕ_{12} denotes the angle XOY, etc.

In the n-dimensional space just mentioned we introduce a new co-ordinate system with the same origin O and in close relation to the points X, Y, and Z (Fig. 2): the x_1-axis is chosen along OX, the x_2-axis perpendicular to OX in the plane OXY, and the x_3-axis perpendicular to the plane OXY in the three-dimensional space $OXYZ$. The remaining $n-3$ axes may be chosen arbitrarily (subject only to the conditions of orthogonality).

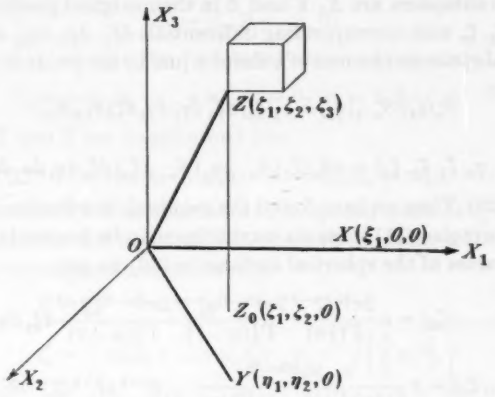


Fig. 2.

The new co-ordinates of the points X, Y and Z are denoted as in Fig. 2 (where only the first three are shown, the remaining being zero). Further we may assume that the change of co-ordinates is carried out in such a manner that $\xi_1 > 0$, $\eta_2 > 0$ and $\zeta_3 > 0$.

The l's being invariant by change of co-ordinates and hence expressed in the same manner by the new and the old co-ordinates, we have

$$\left.\begin{aligned} l_{11} = \xi_1^2, \quad l_{12} = \xi_1\eta_1, \qquad & l_{13} = \xi_1\zeta_1, \\ l_{22} = \eta_1^2 + \eta_2^2, \quad & l_{23} = \eta_1\zeta_1 + \eta_2\zeta_2, \\ & l_{33} = \zeta_1^2 + \zeta_2^2 + \zeta_3^2. \end{aligned}\right\} \tag{31}$$

In relation to any set of points X, Y, Z in the n-dimensional space we may introduce a new co-ordinate system in the above manner and determine the six characteristic quantities $\xi_1, \eta_1, \eta_2, \zeta_1, \zeta_2, \zeta_3$. To begin with the distribution of these six quantities is sought.

With that in view we consider such sets of points X, Y, Z, for which the six quantities have fixed values and where, moreover, the points X and Y are fixed. The projection $Z_0 = (\zeta_1, \zeta_2, 0)$ of Z on the plane OXY will also be fixed, and the distance of Z from the plane has the constant value ζ_3. The point Z, therefore, will be able to move on the surface of an $(n-2)$-dimensional sphere of centre Z_0 and radius ζ_3, situated in an $(n-2)$-dimensional space perpendicular to the plane OXY in Z_0. To Z we attach (see Fig. 2) a small element of

volume $d\zeta_1 d\zeta_2 d\zeta_3$. When Z moves on its spherical surface this element of volume will describe an n-dimensional volume

$$S_{n-2}(\zeta_3)\, d\zeta_1 d\zeta_2 d\zeta_3. \tag{32}$$

Next we consider such sets of points X, Y, where ξ_1, η_1 and η_2 have fixed values, and where moreover X is fixed. The point Y will then be able to move on an $(n-1)$-dimensional sphere of radius η_2 and centre in the projection of Y on OX; a small element of area $d\eta_1 d\eta_2$ attached to Y will thereby describe a volume

$$S_{n-1}(\eta_2)\, d\eta_1 d\eta_2. \tag{33}$$

Finally, when X moves in such a manner that the value of ξ_1 is fixed, then a line-element $d\xi_1$ attached to X will describe an element of volume

$$S_n(\xi_1)\, d\xi_1. \tag{34}$$

We now return to the point M in (30), the projections of which on three mutually orthogonal, n-dimensional subspaces are X, Y and Z in their original positions. To a given set of values $\xi_1, \eta_1, \eta_2, \zeta_1, \zeta_2, \zeta_3$ with corresponding differentials $d\xi_1, d\eta_1, d\eta_2, d\zeta_1, d\zeta_2, d\zeta_3$ we obtain in the $3n$-dimensional space an element of volume equal to the product of the three elements (32), (33) and (34), i.e.

$$S_n(\xi_1)\, S_{n-1}(\eta_2)\, S_{n-2}(\zeta_3)\, d\xi_1 d\eta_1 d\eta_2 d\zeta_1 d\zeta_2 d\zeta_3. \tag{35}$$

Hence we have

$$dP\{\xi_1, \eta_1, \eta_2, \zeta_1, \zeta_2, \zeta_3\} = \kappa S_n(\xi_1)\, S_{n-1}(\eta_2)\, S_{n-2}(\zeta_3)\, d\xi_1 d\eta_1 d\eta_2 d\zeta_1 d\zeta_2 d\zeta_3, \tag{36}$$

where κ is given by (29). Thus we have found the required distribution of $\xi_1, \eta_1, \eta_2, \zeta_1, \zeta_2, \zeta_3$, as by means of the formulae (31) these six quantities may be inserted in κ.

Substituting the values of the spherical surfaces in (36) we get

$$dP\{\xi_1, \eta_1, \dots, \zeta_3\} = \kappa \frac{2\pi^{\frac{1}{2}n} \xi_1^{n-1}}{\Gamma(\frac{1}{2}n)} \frac{2\pi^{\frac{1}{2}(n-1)} \eta_2^{n-2}}{\Gamma\frac{1}{2}(n-1)} \frac{2\pi^{\frac{1}{2}(n-2)} \zeta_3^{n-3}}{\Gamma\frac{1}{2}(n-2)} d\xi_1 d\eta_1 \dots d\zeta_3,$$

thus

$$dP\{\xi_1, \eta_1, \dots, \zeta_3\} = \kappa \frac{8\pi^{\frac{1}{2}(3n-3)}}{\Gamma(\frac{1}{2}n)\, \Gamma\frac{1}{2}(n-1)\, \Gamma\frac{1}{2}(n-2)} \xi_1^{n-1} \eta_2^{n-2} \zeta_3^{n-3} d\xi_1 d\eta_1 \dots d\zeta_3. \tag{37}$$

We shall now find the distribution of the l's introduced in (27). The jacobian of the l's in relation to $\xi_1, \eta_1, \dots, \zeta_3$ may be split up into a product of three and becomes

$$\frac{dl_{11}}{d\xi_1} \frac{\partial(l_{12}, l_{22})}{\partial(\eta_1, \eta_2)} \frac{\partial(l_{13}, l_{23}, l_{33})}{\partial(\zeta_1, \zeta_2, \zeta_3)} = 2\xi_1 \begin{vmatrix} \xi_1 & 0 \\ 2\eta_1 & 2\eta_2 \end{vmatrix} \begin{vmatrix} \xi_1 & 0 & 0 \\ \eta_1 & \eta_2 & 0 \\ 2\zeta_1 & 2\zeta_2 & 2\zeta_3 \end{vmatrix} = 8\xi_1^3 \eta_2^2 \zeta_3.$$

Hence we get

$$dP\{l_{11}, l_{12}, \dots, l_{33}\} = \kappa \frac{\pi^{\frac{1}{2}(3n-3)}}{\Gamma(\frac{1}{2}n)\, \Gamma\frac{1}{2}(n-1)\, \Gamma\frac{1}{2}(n-2)} (\xi_1 \eta_2 \zeta_3)^{n-4} dl_{11} dl_{12} \dots dl_{33}. \tag{38}$$

Moreover, by means of the rule of multiplying determinants we have

$$\begin{vmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{vmatrix} = \begin{vmatrix} \xi_1 & 0 & 0 \\ \eta_1 & \eta_2 & 0 \\ \zeta_1 & \zeta_2 & \zeta_3 \end{vmatrix}^2 = (\xi_1 \eta_2 \zeta_3)^2,$$

whence (38) becomes

$$dP\{l_{11}, l_{12}, \dots, l_{33}\} = \kappa \frac{\pi^{\frac{1}{2}(3n-3)}}{\Gamma(\frac{1}{2}n)\, \Gamma\frac{1}{2}(n-1)\, \Gamma\frac{1}{2}(n-2)} \begin{vmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{vmatrix}^{\frac{1}{2}(n-4)} dl_{11} dl_{12} \dots dl_{33}. \tag{39}$$

This is the required distribution of the l's, as κ is expressed by the l's in (29).

We shall now repeat the proof in a manner which will not involve any geometrical aids. To do this, we proceed once more from (26) and introduce the l's by (27). Furthermore, we introduce X, Y and Z by (28); then

$$dP\{X, Y, Z\} = \kappa dX\,dY\,dZ, \tag{40}$$

where κ is determined by (29). We may for the sake of convenience continue to use the words 'point' and 'co-ordinate'.

We consider an orthogonal transformation

$$\left.\begin{aligned} t_1' &= \frac{1}{\xi_1}(x_1 t_1 + \ldots + x_n t_n) \quad \xi_1 = \sqrt{\left(\sum_1^n x_i^2\right)}, \\ t_i' &= \alpha_{i1} t_1 + \ldots + \alpha_{in} t_n \qquad i = 2, 3, \ldots, n, \end{aligned}\right\} \tag{I}$$

carrying $(t_1, t_2, \ldots, t_n)$ into $(t_1', t_2', \ldots, t_n')$. Applied to the point X it gives

$$X' = (\xi_1, 0, 0, \ldots, 0),$$

all the co-ordinates of X' except the first being equal to zero owing to the conditions of orthogonality

$$\alpha_{i1} x_1 + \alpha_{i2} x_2 + \ldots + \alpha_{in} x_n = 0 \quad (i = 2, 3, \ldots, n).$$

By (I) the points Y and Z are transformed into

$$Y' = (\eta_1, y_2', \ldots, y_n'), \quad Z' = (\zeta_1, z_2', \ldots, z_n'),$$

where the co-ordinates need not be calculated. The special symbols η_1 and ζ_1 are due to the fact that these co-ordinates will not change by the two following transformations.

Next we use a new orthogonal transformation

$$\left.\begin{aligned} t_1'' &= t_1', \\ t_2'' &= \frac{1}{\eta_2}(y_2' t_2' + \ldots + y_n' t_n') \quad \eta_2 = \sqrt{\left(\sum_2^n y_i'^2\right)}, \\ t_i'' &= \beta_{i2} t_2' + \ldots + \beta_{in} t_n' \qquad i = 3, 4, \ldots, n. \end{aligned}\right\} \tag{II}$$

This transformation does not change X', but carries Y' and Z' into

$$Y'' = (\eta_1, \eta_2, 0, \ldots, 0), \quad Z'' = (\zeta_1, \zeta_2, z_3'', \ldots, z_n'').$$

η_1 and ζ_1 remain unchanged as noted above. The last $n-1$ co-ordinates of Z'' need not be calculated. The special symbol ζ_2 is due to the fact that this quantity is unchanged by the following transformation.

Finally, we use a third orthogonal transformation

$$\left.\begin{aligned} t_1''' &= t_1'', \\ t_2''' &= t_2'', \\ t_3''' &= \frac{1}{\zeta_3}(z_3'' t_3'' + \ldots + z_n'' t_n'') \quad \zeta_3 = \sqrt{\left(\sum_3^n z_i''^2\right)}, \\ t_i''' &= \gamma_{i3} t_3'' + \ldots + \gamma_{in} t_n'' \qquad i = 4, 5, \ldots, n. \end{aligned}\right\} \tag{III}$$

This transformation changes neither X' nor Y'', but carries Z'' into

$$Z''' = (\zeta_1, \zeta_2, \zeta_3, 0, \ldots, 0).$$

Thus we have once more introduced the quantities ξ_1, η_1, η_2, ζ_1, ζ_2, ζ_3, known from the geometrical treatment, and as the l's are invariant by orthogonal transformations we again have the equations (31).

We now return to (40) and apply some transformations of variables:

(1) For Y and Z we introduce Y' and Z', respectively, whereas X is left unchanged. As the coefficients of (I) depend only on X, the jacobians are equal to ± 1, and (40) is transformed into

$$dP\{X, \eta_1, Y_1', \zeta_1, Z_1'\} = \kappa dX\, d\eta_1 dY_1' d\zeta_1 dZ_1', \tag{41}$$

where $Y_1' = (y_2', \ldots, y_n')$ and $Z_1' = (z_2', \ldots, z_n')$.

(2) For Z_1' we introduce the last $n-1$ co-ordinates from Z'', while X, η_1, and Y_1' as well as ζ_1 are unchanged. As the coefficients of (II) depend only on Y_1' the jacobian as above is equal to ± 1, and (41) becomes

$$dP\{X, \eta_1, Y_1', \zeta_1, \zeta_2, Z_2''\} = \kappa dX\, d\eta_1 dY_1' d\zeta_1 d\zeta_2 dZ_2'', \tag{42}$$

where $Z_2'' = (z_3'', \ldots, z_n'')$.

According to (31) κ may be expressed by the variables $\xi_1, \eta_1, \eta_2, \zeta_1, \zeta_2, \zeta_3$ only. Hence, as in example 1, instead of X we may introduce $\xi_1 = \sqrt{\left(\sum_1^n x_i^2\right)}$ together with $U = (u_1, u_2, \ldots, u_{n-1})$, for Y_1' similarly $\eta_2 = \sqrt{\left(\sum_2^n y_i'^2\right)}$ together with $V = (v_1, v_2, \ldots, v_{n-2})$ and finally for Z_2'', $\zeta_3 = \sqrt{\left(\sum_3^n z_i''^2\right)}$ and $W = (w_1, w_2, \ldots, w_{n-3})$ as new variables. Eliminating U, V and W by integration we get from (42)

$$dP\{\xi_1, \eta_1, \eta_2, \zeta_1, \zeta_2, \zeta_3\} = \kappa S_n(\xi_1)\, S_{n-1}(\eta_2)\, S_{n-2}(\zeta_3)\, d\xi_1 d\eta_1 d\eta_2 d\zeta_1 d\zeta_2 d\zeta_3.$$

Thus we have once more proved (36) and may as before obtain (37) as well as (39).

Just as we have seen in example 2 that $\sum_1^n (x_i - \bar{x})^2$ is distributed as $\sum_1^n x_i^2$, but with $n-1$ instead of n, it may be shown that the variables

$$l_{11} = \sum_1^n (x_i - \bar{x})^2, \quad l_{12} = \sum_1^n (x_i - \bar{x})(y_i - \bar{y}), \quad \ldots, \quad l_{33} = \sum_1^n (z_i - \bar{z})^2$$

are distributed as the l's in (27), but with $n-1$ instead of n, $\bar{x}$, $\bar{y}$ and $\bar{z}$ being the means of x_i, y_i and z_i respectively. The proof is omitted here.

Finally, it should be noted that the constants of the distributions dealt with in the above all enter automatically and need not be determined by a special process. The common source of all these constants is formula (4), giving the volume of the n-dimensional sphere.

REFERENCES

FISHER, R. A. (1915). *Biometrika*, **10**, 507.
FISHER, R. A. (1925). *Metron*, **5**, 90.
WISHART, J. (1928). *Biometrika*, **20 A**, 32.

PROOFS OF THE DISTRIBUTION LAW OF THE SECOND ORDER MOMENT STATISTICS

By JOHN WISHART, *School of Agriculture, Cambridge*

The foregoing paper by Prof. Fog (1948) calls to mind the different occasions on which proofs have been derived, by various methods, of the result first given twenty years ago by the present author (Wishart, 1928). It may be useful at this stage to catalogue these proofs, so far as known to the writer, and to comment on the contrasting methods employed.

The method first used was a direct extension of the geometrical method of approach used by R. A. Fisher (1915), and was worked out in full for three variates, then extended to the general case by the use of quadratic co-ordinates. At the end of 1933 Prof. Mahalanobis sent the author a somewhat fuller proof on the same lines, which was published some years later (Mahalanobis, Bose & Roy, 1937) as part of a long study of the normalization of statistical variates and the use of rectangular co-ordinates in the theory of sampling distributions. A proof on entirely different lines had been published before the communication referred to from Prof. Malahanobis was received (Wishart & Bartlett, 1933). This depended upon working out the characteristic function of the distribution and then using a generalization of the Fourier integral theorem to express the result as a multiple integral, detailed consideration to which was given by Ingham (1933). Special cases of this method, and of the resulting integrals, had previously been considered for two variables by Romanovsky (1925).

Part of the difficulty, as in all similar problems in the theory of sampling distributions, resides in the fact that the practical problem requires the moment statistics to be calculated from the sample means. This general feature, which is allowed for by writing $n-1$ for n, where n is the size of the sample, may be separated from the fundamental distribution problem, which is that of the simultaneous variation of the $\frac{1}{2}k(k+1)$ sums of squares and products of k variate values $x_i (i = 1, 2, \ldots k)$, which may be taken to follow a multivariate normal distribution with zero means. This was recognized by P. L. Hsu (1939), who obtained a proof by the method of mathematical induction which has the merit of being short. Madow (1938) deduced the distribution by generalizing from Hotelling's joint distribution of sample correlation coefficients. Of the text-book proofs those of Wilks (1943) and Kendall (1946) are geometrical, and that of Cramér (1946) determines the characteristic function. Cramér also refers to a paper by Simonsen (1944, 1945).

The distribution is a direct generalization of the χ^2 distribution to the case of vectors with a number k (>1) of components. Reference should also be made to Bartlett (1933) who deduced the corresponding partial distribution when a number of variates, about whose distribution nothing need be assumed, are eliminated, and to Anderson & Girshick (1944; Anderson, 1946) who deduced cases of the corresponding non-central distribution analogous to the similar problem in the case of χ^2. Recently Elfving (1947) has used matrix methods to deduce Bartlett's decomposition theorem (Bartlett, 1933) and states that the general distribution can be proved in this way, although a formal proof is not given.

Prof. Fog's proof, which provides an interesting parallelism between the geometrical and analytical approaches to the problem, is only given for the case of three variates, but in contradistinction to former proofs his steps would seem to be such as to justify his statement that the generalization to an arbitrary value of k should cause no difficulties. It may not be out of place here to remark that the methods used by him in Examples 1 and 2 may be

used to give a geometrical interpretation to certain special orthogonal transformations which have been used in deriving tests of significance. It is true that the variables U need not be particularly specified, but it helps the elementary student if, for example, $\sum_1^n (x-\bar{x})^2$ and, in the regression problem, $\sum_1^n (y-a-b(x-\bar{x}))^2$ can be directly expressed in a convenient way as the sums of squares of $n-1$ and $n-2$ independent normal variates respectively. A recent expository paper (Wishart, 1947) deduced all the results required for the proof of the variance ratio distribution by using the orthogonal transformation

$$u_i = -\sqrt{\left(\frac{i}{i+1}\right)}(x_{i+1}-\bar{x}_i) \quad (i = 1, 2, \ldots, n-1),$$
$$u_n = \sqrt{n}.\bar{x}_n,$$

where $x_1, x_2, \ldots, x_n$ are normal $(0, 1)$ and $\bar{x}_i = \sum_{r=1}^{i} (x_r)/i$. Then $\sum_1^n (x_i-\bar{x}_n)^2 = \sum_1^{n-1} (u_i^2)$. Similarly, for a regression relationship of the form $Y = bx$ we can make the transformation (Vajda 1945):

$$v_i = -\sqrt{\left(\frac{\Sigma_i}{\Sigma_{i+1}}\right)}(y_{i+1}-b_i x_{i+1}) \quad (i = 1, 2, \ldots, n-1),$$
$$v_n = \sqrt{\Sigma_n}.b_n,$$

where $y_1, y_2, \ldots, y_n$ are normal $(0, 1)$, the x_i can be regarded as fixed, and

$$b_i = \sum_{r=1}^{i} (x_r y_r)/\Sigma_i, \quad \Sigma_i = \sum_{r=1}^{i} (x_r^2).$$

Then
$$\sum_1^n (y_i - b_n x_i)^2 = \sum_1^{n-1} (v_i^2).$$

In both cases (the first is a particular case of the second, in which x_i is put equal to unity for all i), the figure which applies is Fig. 1 of Prof. Fog's paper. $OQ^2 = q_1^2$ represents $\Sigma(x-\bar{x})^2$ in the one case and $\Sigma(y-bx)^2$ in the other, since the linear relations are $\Sigma(x/\sqrt{n}) = \sqrt{n}.\bar{x}$ and $\Sigma(yx/\sqrt{\Sigma_n}) = \sqrt{\Sigma_n}.b$ respectively, and this quantity is therefore the sum of the squares of $(n-1)$ perpendicular distances from Q on to $(n-1)$ hyperplanes through the origin orthogonal to α, and mutually orthogonal among themselves. Any one of an infinity of planes can be chosen for the first, and this corresponds to the statement that the variables U need not be specified. In our first illustration the plane is chosen as $x_1 = x_2$, followed by $x_1 + x_2 = 2x_3$, etc., whereas in the second illustration we start with $x_2 y_1 = x_1 y_2$ (where y_1 and y_2 are the variables, and the x's are constants) and follow up with

$$x_3(x_1 y_1 + x_2 y_2) = y_3(x_1^2 + x_2^2), \text{ etc.}$$

Finally, the distributions required are easily deduced from the fact that the joint probability, after transformation, breaks up into the product of independent parts.

The more general regression relationship $Y = a + b(x-\bar{x})$ is dealt with by applying the first transformation to both the x's and y's, reducing the sums of n squares and products to the sums of $(n-1)$ squares and products of new variables u and u'. To u' is now applied the second transformation (using u) to a variable v', in which i goes from 1 to $n-2$, and

$$v'_{n-1} = \sqrt{\Sigma_{n-1}}.b_{n-1},$$

where
$$\Sigma_{n-1} = \sum_{r=1}^{n-1} (u_r^2), \quad b_{n-1} = \sum_{r=1}^{n-1} (u_r u')/\Sigma_{n-1}.$$

We then have
$$\sum_1^n (y-Y)^2 = \sum_1^{n-2} (v_i'^2)$$

(Wishart, 1948). This represents a particular application of Fisher's general theorem (1925).

Geometrically, the position is as represented in the diagram, which is a further development of Prof. Fog's Fig. 1. The first relation $\sum_1^n (y/\sqrt{n}) = \sqrt{n}.\bar{y}$ projects the point Y on to plane I through the origin in the point $Q(y_1-\bar{y}, \ldots, y_n-\bar{y})$, and $OY^2 = z_1^2+q_1^2$, where $z_1 = \sqrt{n}.\bar{y}$ and $q_1 = \sqrt{[\Sigma(y-\bar{y})^2]}$; the second relation $\sum_1^{n-1} (uu'/\sqrt{\Sigma_{n-1}}) = \sqrt{\Sigma_{n-1}}.b$ projects Q on to a plane II through the origin (orthogonal to plane I) in the point $Q'(y_1-Y_1, \ldots, y_n-Y_n)$, and $q_1^2 = z_2^2+q_2^2$, where $z_2 = \sqrt{\Sigma_{n-1}}.b$, and $q_2 = \sqrt{[\Sigma(y-Y)^2]}$. Q' is confined to the $(n-2)$ space determined by the intersection of planes I and II. Altogether $OY^2 = z_1^2+z_2^2+q_2^2$, and q_2^2 can then be determined as the sum of the squares of $(n-2)$ perpendicular distances from Q' on to $(n-2)$ other hyperplanes through the origin orthogonal to plane II, and mutually orthogonal among themselves. Again an infinity of planes can be chosen for the first, and the one we take is $u_2u_1' = u_1u_2'$, which can if desired be expressed in terms of the original variables x and y as $(x_2-x_3)y_1 + (x_3-x_1)y_2 + (x_1-x_2)y_3 = 0$. The mutual independence of the distributions of $\bar{y}\sqrt{n}$, $b\sqrt{[\Sigma(x-\bar{x})^2]}$ and $\Sigma(y-Y)^2$, and their separate distribution functions, then follow.

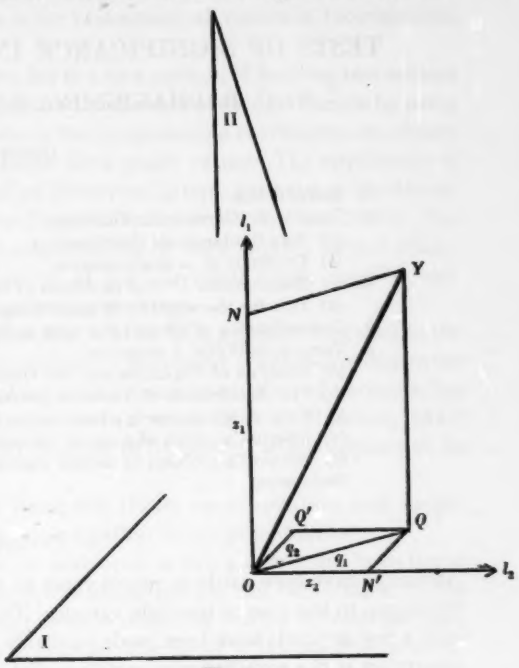


Fig. 1.

REFERENCES

ANDERSON, T. W. (1946). *Ann. Math. Statist.* **17**, 409.
ANDERSON, T. W. & GIRSHICK, M. A. (1944). *Ann. Math. Statist.* **15**, 345.
BARTLETT, M. S. (1933). *Proc. Roy. Soc. Edinb.* **53**, 260.
CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.
ELFVING, G. (1947). *Skand. AktuarTidskr.* **30**, 56.
FISHER, R. A. (1915). *Biometrika*, **10**, 507.
FISHER, R. A. (1925). *Metron*, **5**, no. 3, p. 90.
FOG, D. (1948). *Biometrika*, **35**, 46.
HSU, P. L. (1939). *Proc. Camb. Phil. Soc.* **35**, 336.
INGHAM, A. E. (1933). *Proc. Camb. Phil. Soc.* **29**, 270.
KENDALL, M. G. (1946). *The Advanced Theory of Statistics*, vol. II. London: Griffin and Co.
MADOW, W. G. (1938). *Trans. Amer. Math. Soc.* **44**, 454.
MAHALANOBIS, P. C., BOSE, R. C. & ROY, S. N. (1937). *Sankhyā*, **3**, 1.
ROMANOVSKY, V. (1925). *Metron*, **5**, no. 4, p. 3.
SIMONSEN, W. (1944, 1945). *Skand. AktuarTidskr.* **27**, 235; **28**, 20.
VAJDA, S. (1945). *Ann. Math. Statist.* **16**, 381.
WILKS, S. S. (1943). *Mathematical Statistics*. Princeton University Press.
WISHART, J. (1928). *Biometrika*, **20A**, 32.
WISHART, J. (1947). *J. Inst. Actu. Stud. Soc.* **7**, 98.
WISHART, J. (1948). *J. Inst. Actu. Stud. Soc.* **8**, 38.
WISHART, J. & BARTLETT, M. S. (1933). *Proc. Camb. Phil. Soc.* **29**, 260.

TESTS OF SIGNIFICANCE IN MULTIVARIATE ANALYSIS

By C. RADHAKRISHNA RAO, *King's College, Cambridge*

CONTENTS

	PAGE
1. Introduction	58
2. Tests with discriminant functions	
(*a*) Two fundamental distributions	59
(*b*) Problems of a single sample	60
(*c*) Mahalanobis' D^2 and problems of two samples	63
(*d*) Test for the equality of discriminant functions	66
3. Generalization of D^2 and the large sample theory for several groups	66
4. Tests with Wilks' Λ criterion	
(*a*) Analysis of dispersion and the theoretical aspects of the Λ criterion	67
(*b*) The distribution of Λ and its practical use	69
(*c*) Test of differences in mean values for several populations	71
(*d*) Internal analysis of a set of variates	73
(*e*) Barnard's problem of secular variations in skull characters	74
References	78

1. Introduction

Attempts have been made in recent years to generalize the univariate analysis of variance technique to the case of multiple variates. The extension of the theory has been slow and only a few methods have been made available for practical use. The starting-point of these researches is the simultaneous sampling distribution of the variances and covariances in samples from a multivariate normal population given by Wishart in 1928. A few years later Hotelling (1931) found the distribution of a quantity T which is a natural extension of Student's distribution to a sample from a multivariate normal population.

Wilks (1932) following the *likelihood ratio method* (Neyman & Pearson, 1928, 1931; Pearson & Neyman, 1930) obtained suitable generalizations in the analysis of variance applicable to several variables. The statistic Λ proposed by him has been found useful in a variety of problems. Bartlett (1934) applied it for testing the significance of treatments with respect to two variables in a varietal trial and indicated its general use in multivariate tests of significance. Wilks (1935) and Hotelling (1935) found it useful in testing the independence of several groups of variates. Recently Plackett (1947) provided an exact test for judging the equality of variances and covariances of various populations, with the use of this criterion. Wilks' statistic supplied some of the basic tests in multivariate analysis but the problem of tabulation has not been tackled except in some limited cases (Wald & Brookner, 1941). A very useful approximation has been suggested by Bartlett (1938) who further demonstrated its use in a paper on 'Multivariate Analysis' read before the Royal Statistical Society in May 1947.

A new line of research was initiated by Fisher (1936) with his introduction of the discriminant function analysis. It has been shown (Barnard, 1935; Martin, 1935; Fairfield Smith, 1936) that a set of multiple measurements may be used to provide a discriminant function *linear in the observations* having the property that, better than *any other linear function*, it will discriminate between any two chosen classes such as taxonomic species, the

two sexes and so on. It has also been shown (Welch, 1939; Rao, 1947) that the linear discriminant function chosen as stated above is the best among all classes of functions not necessarily linear in the set of observations.

The introduction of the discriminant function led to a new method of deriving test criteria suitable for multiple variates. The problem is reduced to the case of a single variate by using a linear compound of the several variables, where the compounding coefficients are chosen so as to maximize the value of a statistic suitable for a single variate. The application of this method to test the differences in mean values for several groups gave rise to the theory of canonical roots of determinantal equations (Roy, 1939; Fisher, 1939; Hsu, 1939*b*). The distribution of the individual roots and the exact nature of tests require further study. Wilks' statistic, which is a symmetric function of the canonical roots, may be considered as providing an overall test of the hypothesis concerned.

The object of the present paper is twofold. The first is to develop a unified approach to the problem of tests of significance in multivariate analysis. The concept of *analysis of dispersion* explained in § 4 (*a*), which is a natural extension of the univariate analysis of variance, has been found useful in discussing multivariate problems. In a recent abstract, Hotelling (1947) developed a method of splitting a generalized measure of dispersion which appears to be different from the method proposed here.

The second is to examine the nature of Bartlett's (1938) approximation and supply appropriate methods for cases needing the exact evaluation of the probabilities.

In presenting the various tests of significance developed in this paper it has been found convenient to consider the problems arising out of a single sample and two samples in the first stage. They depend on simple tests of significance requiring the use of variance ratio tables alone and are of very great importance in practice. The use of Wilks' statistic in multivariate analysis involving more than two samples is considered in the second stage. A number of examples have been worked out to explain the computational procedure.

2. TESTS WITH DISCRIMINANT FUNCTIONS

(*a*) *Two fundamental distributions*

The method of discriminant functions to derive test criteria has been found extremely useful in multivariate analysis. The problem is reduced to the case of a single variable by choosing a linear compound of the variables and constructing a statistic suitable for the case of a single variate. The maximized value of this statistic obtained by a suitable choice of the compounding coefficients is taken as the appropriate test criterion. The distribution of the statistics thus derived in problems involving a single sample and two samples depend on the two fundamental distributions considered below.

Let (w_{ij}), $i, j = 1, 2, \ldots, p$ be the matrix giving the estimates, on n degrees of freedom, of the elements in the dispersion matrix (α_{ij}) of p normally correlated variables. The definition of w_{ij} implies that it has been calculated from a certain sum of products by dividing by the appropriate degrees of freedom. Let $d_1, d_2, \ldots, d_p$ be p normal variates with the same dispersion matrix (α_{ij}) but distributed independently of the w_{ij}'s. Considering only the first r variables $d_1, \ldots, d_r$ the statistic V_r is defined as

$$nV_r = \sum_{i=1}^{r} \sum_{j=1}^{r} w_r^{ij} d_i d_j, \qquad (2{\cdot}1)$$

where (w_r^{ij}) is the matrix reciprocal to (w_{ij}), $i, j = 1, 2, \ldots, r$. An alternative method of calculating this statistic is provided by the equation

$$1 + nV_r = \frac{|w_{ij} + d_i d_j|}{|w_{ij}|} \quad (i, j = 1, 2, \ldots, r).$$

It has been shown by Hotelling (1931) that, when the w_{ij} follow Wishart's distribution (Wishart, 1928) with n degrees of freedom and $E(d_1) = E(d_2) = \ldots = E(d_r) = 0$, the statistic $V_r(n+1-r)/r$ can be used as a variance ratio with r and $(n+1-r)$ degrees of freedom.

The author has shown elsewhere (Rao, 1946*b*) that, if $d_{r+1}, \ldots, d_p$ are distributed independently of $d_1, \ldots, d_r$ and $E(d_{r+1}) = \ldots = E(d_p) = 0$, $E(d_i)$ being not necessarily zero when $i = 1, 2, \ldots, r$, the statistic

$$U = \frac{n+1-p}{p-r}\left\{\frac{1+V_p}{1+V_r} - 1\right\} \tag{2·2}$$

can be used as a variance ratio with $(p-r)$ and $(n+1-p)$ degrees of freedom. The statistic V_p is calculated from the formula (2·1) by using all the p variates.

All the tests of significance considered in this section depend on the use of the statistics defined in (2·1) and (2·2).

(*b*) *Problems of a single sample*

Student's test connected with pairs of observations admits generalization in two directions.

The first is to test whether the means of p correlated variables are the same on the basis of a sample of size N from a p-variate population. When the test shows differences in mean values, there arises the question of deciding whether an assigned contrast involving the p variates differs from the best contrast as determined from the data.

If $x_{1i}, x_{2i}, \ldots, x_{pi}$ are the observations on the ith individual, then they may be replaced by a linear compound $z_i = l_1 x_{1i} + \ldots + l_p x_{pi}$ where the l's satisfy the condition $l_1 + \ldots + l_p = 0$. The problem of determining the best contrast reduces to that of determining the compounding coefficients $l_1, \ldots, l_p$ such that the ratio of mean z to standard deviation of z is a maximum. An alternative method which has some practical advantage is as follows.

By arbitrary choice of constants one can construct $(p-1)$ linear combinations of the variables $x_1, \ldots, x_p$,

$$y_j = m_{1j}x_1 + \ldots + m_{pj}x_p,$$

such that $\sum_i m_{ij} = 0$ for $j = 1, 2, \ldots, (p-1)$. Choosing a linear compound of the x's with coefficients adding to zero is the same as choosing a linear compound of the y's without any restriction on the compounding coefficients. If the linear compound is

$$\lambda_1 y_1 + \lambda_2 y_2 + \ldots + \lambda_{p-1} y_{p-1},$$

then the quantity to be maximized is

$$v = \frac{(\lambda_1 \bar{y}_1 + \ldots + \lambda_{p-1}\bar{y}_{p-1})^2}{\Sigma\Sigma \lambda_i \lambda_j w_{ij}},$$

where

$$w_{ij} = \frac{1}{N-1}\sum_{r=1}^{N}(y_{ir} - \bar{y}_i)(y_{jr} - \bar{y}_j).$$

Observing that only the ratios of λ's are uniquely determinable the equations giving λ's may be written as $\lambda_1 w_{1i} + \ldots + \lambda_{p-1} w_{p-1i} = \bar{y}_i \quad (i = 1, 2, \ldots, (p-1))$,

with the solution as $\lambda_i = w^{1i}\bar{y}_1 + \ldots + w^{p-1i}\bar{y}_{p-1} \quad (i = 1, 2, \ldots, (p-1))$,

where the matrix (w^{ij}) is reciprocal to (w_{ij}). This supplies the best linear compound of the y's which on transformation to the x's gives the best contrast determinable from the data.

The maximum value of v is given by

$$\Sigma\lambda_i\bar{y}_i = \Sigma\Sigma w^{ij}\bar{y}_i\bar{y}_j.$$

If $V_{p-1} = N(\Sigma\Sigma w^{ij}\bar{y}_i\bar{y}_j)/(N-1)$, then, on the hypothesis that all the x's have the same mean value, the conditions required for the use of the statistic (2·1) are satisfied so that

$$V_{p-1}(N-p+1)/(p-1)$$

can be used as the variance ratio with $(p-1)$ and $(N-p+1)$ degrees of freedom to test the above hypothesis.

The statistic V_{p-1} is invariant for all sets of coefficients chosen to construct the y's from the x's so that in any practical problem either conveniently or conventionally chosen linear contrasts of the x's may be used to define the y's.

The test used above is essentially the one given by Hotelling (1931). The point of interest is to show how the test is derivable by the method of discriminant functions involving some restrictions on the compounding coefficients. Also the author is not aware whether the use of Hotelling's T in testing the equality of means of p correlated variables has been explicitly mentioned anywhere.

To test whether the best contrast as determined from the data is in agreement with an assigned contrast $\xi_1x_1+\ldots+\xi_px_p$ or $\eta_1y_1+\ldots+\eta_{p-1}y_{p-1}$ in terms of the y's one may proceed as follows.

The appropriate statistic for testing the significance of the assigned contrast is

$$V_1 = \frac{N(\eta_1\bar{y}_1+\ldots+\eta_{p-1}\bar{y}_{p-1})^2}{(N-1)(\Sigma\Sigma\eta_i\eta_j w_{ij})},$$

where $V_1(N-1)$ is the variance ratio with 1 and $(N-1)$ degrees of freedom. The appropriate statistic for all the $(p-1)$ contrasts is V_{p-1} as considered before. The hypothesis specifies that all contrasts orthogonal to the assigned one have zero mean so that the conditions for the use of the statistic (2·2) are satisfied. Hence

$$U = \frac{(N-p+1)}{(p-2)}\left\{\frac{1+V_{p-1}}{1+V_1}-1\right\}$$

can be used as the variance ratio with $(p-2)$ and $(N-p+1)$ degrees of freedom to test the above hypothesis.

The above test can be generalized to answer the problem whether a set of k assigned contrasts contain the best contrast. In this case the statistic

$$U = \frac{(N-p+1)}{(p-k-1)}\left\{\frac{1+V_{p-1}}{1+V_k}-1\right\}$$

can be used as the variance ratio with $(p-k-1)$ and $(N-p+1)$ degrees of freedom.

The second generalization of Student's t is concerned with testing, on the basis of a sample of size N from a $2p$-variate population containing the variables $y_1,\ldots,y_{2p}$ whether $E(y_i) = E(y_{i+p})$ for $i = 1, 2, \ldots, p$.

From the $2p$ variates $y_1,\ldots,y_{2p}$ one can construct the p variates $z_i = y_i - y_{i+p}$, $i = 1, 2, \ldots, p$ in which case the problem reduces to that of testing the hypothesis $E(z_i) = 0$, $i = 1, 2, \ldots, p$. The variance ratio with p and $(N-p)$ degrees of freedom to test the above hypothesis is

$$\frac{N(N-p)}{p(N-1)}\Sigma\Sigma w^{ij}\bar{z}_i\bar{z}_j,$$

where (w^{ij}) is reciprocal to (w_{ij}) giving the estimates of the variances and covariances of the z's.

This test may be useful in biometry where the asymmetry of organisms is considered. The sets of variables $y_1, \ldots, y_p$ and $y_{p+1}, \ldots, y_{2p}$ will then correspond to measurements on the right and left sides of an organism.

Example 1. The data of Table 1 consist of weights of cork borings taken by the author from the north (N), east (E), south (S) and west (W) directions of the trunk for 28 trees in a block of plantations. The problem is to test whether the bark deposit varies in thickness and hence in weight in the four directions. It was suggested by Prof. Mahalanobis that the bark deposit is likely to be uniform in N and S directions and also uniform but less in E and W directions, so that N − E − W + S can be taken as the best contrast. This can, however, be tested from the given data as shown below.

Table 1. *Weights of cork borings (in centigrams) in the four directions for* 28 *trees*

N	E	S	W	N	E	S	W
72	66	76	77	91	79	100	75
60	53	66	63	56	68	47	50
56	57	64	58	79	65	70	61
41	29	36	38	81	80	68	58
32	32	35	36	78	55	67	60
30	35	34	26	46	38	37	38
39	39	31	27	39	35	34	37
42	43	31	25	32	30	30	32
37	40	31	25	60	50	67	54
33	29	27	36	35	37	48	39
32	30	34	28	39	36	39	31
63	45	74	63	50	34	37	40
54	46	60	52	43	37	39	50
47	51	52	43	48	54	57	43

It has been found in similar studies that there exists a significant correlation between contrasts such as (N − E) and (S − W) so that the method of fitting constants for the four directions and the individual trees by the method of least squares is not appropriate. The three contrasts arising out of the four weights may, however, be treated as three correlated variables in which case the theory developed above is applicable.

It is interesting to observe that the individual weights in Table 1 are exceedingly asymmetrically distributed.* This does not, however, invalidate the test so long as the contrasts are normally distributed. In fact, the distribution of the individual weights depends on the nature of the plants and the variation between plants. If the above condition is satisfied, it is not necessary that the individual weights should follow any distribution law of the known type. It may be sometimes necessary to make a transformation (such as log, square or cube root) of the variables under consideration to ensure that the contrasts of the transformed variables are symmetrically distributed if the contrasts of the original variables are not so.

As observed earlier the contrasts may be conveniently or conventionally chosen. For the above example one may choose the simple set of contrasts

$$y_1 = \mathrm{N} - \mathrm{E} - \mathrm{W} + \mathrm{S}, \quad y_2 = \mathrm{S} - \mathrm{W}, \quad y_3 = \mathrm{N} - \mathrm{S}.$$

* I am indebted to Prof. E. S. Pearson for drawing my attention to this fact.

The mean values and estimates of variances and covariances based on 27 degrees of freedom for the y's are $\bar{y}_1 = 8{\cdot}8571, \quad \bar{y}_2 = 4{\cdot}5000, \quad \bar{y}_3 = 0{\cdot}8571,$

$$(w_{ij}) = \begin{pmatrix} 128{\cdot}7200 & 61{\cdot}4076 & -21{\cdot}0211 \\ 61{\cdot}4076 & 56{\cdot}9259 & -28{\cdot}2963 \\ -21{\cdot}0211 & -28{\cdot}2963 & 63{\cdot}5344 \end{pmatrix}.$$

The coefficients in the best linear function $\lambda_1 y_1 + \lambda_2 y_2 + \lambda_3 y_3$ are given by the equations

$$\begin{aligned} 128{\cdot}7200\lambda_1 + 61{\cdot}4076\lambda_2 - 21{\cdot}0211\lambda_3 &= 8{\cdot}8571, \\ 61{\cdot}4076\lambda_1 + 56{\cdot}9259\lambda_2 - 28{\cdot}2963\lambda_3 &= 4{\cdot}5000, \\ -21{\cdot}0211\lambda_1 - 28{\cdot}2963\lambda_2 + 63{\cdot}5344\lambda_3 &= 0{\cdot}8571. \end{aligned}$$

Solving, one gets $\lambda_1 = 0{\cdot}05620, \quad \lambda_2 = 0{\cdot}04415, \quad \lambda_3 = 0{\cdot}05174,$

so that the best contrast is

$$\begin{aligned} &\lambda_1(\mathrm{N}-\mathrm{E}-\mathrm{W}+\mathrm{S}) + \lambda_2(\mathrm{S}-\mathrm{W}) + \lambda_3(\mathrm{N}-\mathrm{S}) \\ &\quad = 0{\cdot}10794\mathrm{N} - 0{\cdot}05620\mathrm{E} - 0{\cdot}10035\mathrm{W} + 0{\cdot}04861\mathrm{S}, \end{aligned}$$

or

$$1{\cdot}0794\mathrm{N} - 0{\cdot}5620\mathrm{E} - 1{\cdot}0035\mathrm{W} + 0{\cdot}4861\mathrm{S},$$

obtained by multiplying the coefficients by 10 (arbitrarily).

The statistic for testing the hypothesis of equality of means is

$$\begin{aligned} V_{p-1} &= \frac{N}{N-1}(\lambda_1\bar{y}_1 + \lambda_2\bar{y}_2 + \lambda_3\bar{y}_3) \\ &= \tfrac{28}{27}(0{\cdot}740790) = 0{\cdot}768226. \end{aligned}$$

$$V_{p-1}(N-p+1)/(p-1) = 0{\cdot}768226(28-4+1)/3 = 6{\cdot}4019.$$

The quantity 6·4019 as the variance ratio with 3 and 25 degrees of freedom is significant at 1% level so that the bark deposit cannot be considered uniform in the four directions.

The assigned contrast is represented by y_1 and to test for its significance one can construct the statistic

$$V_1 = \frac{N}{N-1}\frac{\bar{y}_1^2}{w_{11}} = \frac{28}{27}\frac{(8{\cdot}8571)^2}{128{\cdot}7200} = 0{\cdot}632020.$$

The quantity $(N-1)V_1 = 17{\cdot}0645$ as the variance ratio with 1 and 27 degrees of freedom is highly significant.

To test whether the assigned contrast agrees with that estimated from the data, the statistic U defined in (2·2) has to be calculated.

$$U = \frac{N-p+1}{p-2}\left\{\frac{1+V_{p-1}}{1+V_1} - 1\right\} = \frac{25}{2}\left\{\frac{1{\cdot}768226}{1{\cdot}632020} - 1\right\} = 1{\cdot}0431.$$

This value of U as the variance ratio with 2 and 25 degrees of freedom is small so that the evidence supplied by the data is not sufficient to reject the assigned contrast as not the best, although the ratios of the coefficients in the estimated contrast depart considerably from those assigned.

(c) *Mahalanobis' D^2 and problems of two samples*

Let N_1 and N_2 be the sample sizes from two populations π_1 and π_2 characterized by $(p+q)$ variates. The sample means for the ith character are represented by $\bar{x}_{i1}$ and $\bar{x}_{i2}$ for π_1 and π_2 respectively. The estimated value of the covariance is given by

$$(N_1+N_2-2)\,w_{ij} = \sum_{t=1}^{N_1}(x_{i1t}-\bar{x}_{i1})(x_{j1t}-\bar{x}_{j1}) + \sum_{t=1}^{N_2}(x_{i2t}-\bar{x}_{i2})(x_{j2t}-\bar{x}_{j2}).$$

Mahalanobis' (1936) distance between the two populations as estimated from the sample on the basis of the first p characters is

$$D_p^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} w_p^{ij} (\bar{x}_{i1} - \bar{x}_{i2}) (\bar{x}_{j1} - \bar{x}_{j2}),$$

where (w_p^{ij}) is reciprocal to (w_{ij}), $i, j = 1, 2, \ldots, p$. The exact distribution of D_p^2 on the hypothesis specifying real differences in mean values has been given by Bose & Roy (1938). To test the hypothesis specifying no difference in mean values of the p characters for π_1 and π_2 the statistic

$$\frac{N_1 N_2 (N_1 + N_2 - p - 1)}{p(N_1 + N_2)(N_1 + N_2 - 2)} D_p^2$$

can be used as the variance ratio with p and $(N_1 + N_2 - 1 - p)$ degrees of freedom. The appropriate distance function based on $(p+q)$ characters is

$$D_{p+q}^2 = \Sigma\Sigma w_{p+q}^{ij} (\bar{x}_{i1} - \bar{x}_{i2}) (\bar{x}_{j1} - \bar{x}_{j2}).$$

To test whether the q additional characters lead to significant differences between π_1 and π_2, independently of the first p characters, a comparison may be made of the magnitudes of D_p^2 and D_{p+q}^2. One may choose the ratio

$$R = \frac{1 + N_1 N_2 D_{p+q}^2 / (N_1 + N_2)(N_1 + N_2 - 2)}{1 + N_1 N_2 D_p^2 / (N_1 + N_2)(N_1 + N_2 - 2)}$$

which has some theoretical advantage as observed in the next section. The conditions for the use of the statistic defined in (2·2) are satisfied so that

$$U = \frac{N_1 + N_2 - p - q - 1}{q} (R - 1)$$

is the variance ratio with q and $(N_1 + N_2 - p - q - 1)$ degrees of freedom.

It has been shown by Fisher (1938) that the test with Mahalanobis' D^2 is equivalent to testing the difference in the mean values of the discriminant function as estimated from the samples. Apart from tests of significance, the discriminant function is used for the purpose of assigning an individual to its proper class. This leads us to the problem of determining the number and nature of characters used in order that the errors of classification may not be large. It is, however, to be expected that the errors of classification will decrease as the number of characters is increased. On the other hand, since the various characters are correlated with one another the addition of some characters to a basic panel may not reduce or at any rate substantially decrease the errors. In this case, it may not be worth while to increase the number of characters so that the numerical computation involved may not be heavy. A test designed to judge the significance of the reduction in errors of classification by the addition of some new characters is valuable in problems of this nature. Such a test is mathematically equivalent to a test of differences in mean values of the new variables after eliminating the differences in the basic characters. If p is the number of basic characters and q the number of additional characters the test may also be conceived as a comparison of Mahalanobis' distances based on p and $(p+q)$ characters.

If it is desired to test whether an assigned discriminant function $\eta_1 x_1 + \ldots + \eta_p x_p$ of p characters differs from that calculated from the data, then the statistic

$$\frac{N_1 + N_2 - p - 1}{p - 1} \left\{ \frac{1 + N_1 N_2 D_p^2 / (N_1 + N_2)(N_1 + N_2 - 2)}{1 + N_1 N_2 D_1^2 / (N_1 + N_2)(N_1 + N_2 - 2)} - 1 \right\}$$

where $D_1^2 = \Sigma\Sigma \eta_i \eta_j (\bar{x}_{i1} - \bar{x}_{i2}) (\bar{x}_{j1} - \bar{x}_{j2}) / \Sigma\Sigma \eta_i \eta_j w_{ij}$, can be used as the variance ratio with $(p-1)$ and $(N_1 + N_2 - p - 1)$ degrees of freedom (Fisher, 1940; Bartlett, 1939).

Example 2. The following Tables 2 and 3 reproduced from Fisher (1938) give the mean values based on fifty observations each and the covariances based on $50+50-2$ degrees of freedom for four characters in two species of plants *Iris versicolor* and *Iris setosa*.

The solution of the equations

$$\lambda_1 w_{1i}+\ldots+\lambda_4 w_{4i}=d_i \quad (i=1,2,3,4)$$

is obtained as

$$\lambda_1=-3{\cdot}0528, \quad \lambda_2=-18{\cdot}0229, \quad \lambda_3=21{\cdot}7662, \quad \lambda_4=30{\cdot}8442,$$

so that the discriminant function is

$$-3{\cdot}0528x_1-18{\cdot}0229x_2+21{\cdot}7662x_3+30{\cdot}8442x_4.$$

The value of D_4^2 is* $\quad \lambda_1 d_1+\ldots+\lambda_4 d_4=103{\cdot}2335.$

To test for the differences in mean values one can use the statistic

$$\frac{N_1N_2(N_1+N_2-1-4)}{(N_1+N_2)(N_1+N_2-2)}\frac{D_4^2}{4}=\frac{50\times50\times95}{100\times98\times4}\times103{\cdot}2335$$
$$=\tfrac{95}{4}(26{\cdot}3350)=625{\cdot}4515,$$

which as the variance ratio with 4 and 95 degrees of freedom is highly significant.

Table 2. *Observed mean values based on fifty observations each for the two species*

Character	*Iris versicolor*	*Iris setosa*	Difference (d)
Sepal length (x_1)	5·936	5·006	0·930
Sepal width (x_2)	2·770	3·428	−0·658
Petal length (x_3)	4·260	1·462	2·798
Petal width (x_4)	1·326	0·246	1·080

Table 3. *The pooled covariance matrix based on* 98 *degrees of freedom*

	x_1	x_2	x_3	x_4
x_1	0·195340	0·092200	0·099626	0·033055
x_2	0·092200	0·121079	0·047175	0·025251
x_3	0·099626	0·047175	0·125488	0·039586
x_4	0·033055	0·025251	0·039586	0·025106

If only the lengths are considered (i.e. x_1 and x_3) the equations leading to the discriminant function are

$$0{\cdot}195340\mu_1+0{\cdot}099626\mu_2=0{\cdot}930,$$
$$0{\cdot}099626\mu_1+0{\cdot}125488\mu_2=2{\cdot}798,$$

so that $\mu_1=-11{\cdot}1088$, $\mu_2=31{\cdot}1163$ and

$$D_2^2=\mu_1(0{\cdot}930)+\mu_2(2{\cdot}798)=76{\cdot}7322.$$

To answer the question whether the widths (x_2 and x_4) supply independent information the significance of the ratio

$$R=\frac{1+N_1N_2D_4^2/(N_1+N_2)(N_1+N_2-2)}{1+N_1N_2D_2^2/(N_1+N_2)(N_1+N_2-2)}$$
$$=\frac{1+26{\cdot}3350}{1+19{\cdot}5745}=1{\cdot}3286.$$

* This method of calculating D^2 avoids the actual inversion of the matrix (w_{ij}). The λ coefficients obtained by this process can be directly used in the construction of the discriminant function.

has to be tested. The value of the statistic

$$U = (R-1)(N_1+N_2-1-4)/2 = 95(0{\cdot}3286)/2 = 15{\cdot}6085$$

as the variance ratio with 2 and 95 degrees of freedom is significant at 1 % level. This shows that the widths in association with the lengths lead to further discrimination of the species, so that there is a significant reduction in the errors of classification.

(*d*) *Test for the equality of discriminant functions*

If four samples of sizes N_1, N_2, N_3 and N_4 from populations π_1, π_2, π_3 and π_4 are available one can test whether the discriminant functions between π_1, π_2 and π_3, π_4 are significantly different by an extension of the test criterion discussed above. It is a necessary condition of the test that the variances and covariances are identical in the four populations π_1, π_2, π_3 and π_4. No reasonably simple test can be constructed to establish the equivalence of the discriminant functions when this condition is not satisfied.

Let (w_{ij}) be the dispersion matrix based on $(N_1+N_2+N_3+N_4-4)$ degrees of freedom. If $d_1, \ldots, d_p$ are the differences in mean values for π_1 and π_2 and $d'_1, \ldots, d'_p$ are those for π_3 and π_4, the test for equality of discriminant functions is identical with the testing of the hypothesis

$$E(d_i) = E(d'_i) \quad (i = 1, 2, \ldots, p),$$

or

$$E(d_i) = E(-d'_i) \quad (i = 1, 2, \ldots, p).$$

The variance ratios with p and $n = (N_1+N_2+N_3+N_4-3-p)$ degrees of freedom for the two cases are

$$\frac{n}{p}\frac{f(N)}{n+p-1}\Sigma\Sigma w^{ij}(d_i-d'_i)(d_j-d'_j)$$

and

$$\frac{n}{p}\frac{f(N)}{n+p-1}\Sigma\Sigma w^{ij}(d_i+d'_i)(d_j+d'_j)$$

where

$$\frac{1}{f(N)} = \left(\frac{1}{N_1}+\frac{1}{N_2}+\frac{1}{N_3}+\frac{1}{N_4}\right).$$

The equality of discriminant functions is indicated if at least one of the statistics is not significant. Similar tests can be constructed for judging the differences in discriminant functions in parallel samples from two populations or between π_1, π_2 and π_1, π_3.

3. Generalization of D^2 and the large sample theory for several groups

Let there be k multivariate populations $\pi_1, \pi_2, \ldots, \pi_k$ from which samples of sizes $N_1, N_2, \ldots, N_k$ are available for $p+q$ characters. The common covariance matrix assumed to be known or estimated on a large number of degrees of freedom is represented by $(\alpha_{ij}^{(p)})$ for the first p characters and by $(\alpha_{ij}^{(p+q)})$ for all the $(p+q)$ characters. The inverse of $(\alpha_{ij}^{(p)})$ is represented by $(\alpha_{(p)}^{ij})$ and that of $(\alpha_{ij}^{(p+q)})$ by $(\alpha_{(p+q)}^{ij})$. Let $\bar{x}_{i1}, \bar{x}_{i2}, \ldots$ be the mean values of the ith character in the first, second, etc., populations.

It has been shown by Hsu (1939*a*) and Rao (1945) that the statistic,

$$V_{p,k} = \sum_{i,j=1}^{p}\sum\alpha_{(p)}^{ij}\sum_{r=1}^{k}N_r(\bar{x}_{ir}-\bar{x}_{i.})(\bar{x}_{jr}-\bar{x}_{j.}),$$

where $\bar{x}_{i.} = (\Sigma N_r\bar{x}_{ir})/(\Sigma N_r)$, can be used as χ^2 with $p(k-1)$ degrees of freedom to test the hypothesis that the mean values are the same in all the k populations for these p characters. The statistic V_{pk} is a suitable generalization of the Mahalanobis' D^2 in its classical form and its theoretical derivation has been discussed by the author (Rao, 1945).

When this test indicates differences in mean values it is, in some problems, necessary to test whether the observations on q additional characters supply independent information for discrimination. The statistic for testing the differences in means for all the $p+q$ characters is

$$V_{p+q,k} = \sum_{i,j=1}^{p+q} \alpha_{(p+q)}^{ij} \sum_{r=1}^{k} N_r(\bar{x}_{ir} - \bar{x}_{i.})(\bar{x}_{jr} - \bar{x}_{j.}),$$

which can be used as a χ^2 with $(p+q)(k-1)$ degrees of freedom. The $q(k-1)$ additional degrees of freedom bring in the contribution

$$V_{p+q,k} - V_{p,k},$$

and the significance of this difference can be appropriately used to judge the significance of the information supplied by the additional characters. This difference can be used as a χ^2 with $q(k-1)$ degrees of freedom as shown below.

The hypothesis that the new characters do not lead to further discrimination of the populations specifies that any linear function of the $(p+q)$ characters uncorrelated with each of the p characters has the same mean value for all the k populations. There are q such linear functions and treating them as q variables a χ^2 with $q(k-1)$ degrees of freedom can be constructed to test the above hypothesis. The above method of taking the difference is only an alternative way of calculating this χ^2. For $V_{p+q,k}$ calculated from all the $(p+q)$ characters, being invariant under linear transformations of the variables, is equal to $V_{p,k} + \chi^2$ calculated from the p original characters and the q linear functions chosen to be uncorrelated with each of the p characters.

In the above derivation it has been assumed that the variances and covariances are known and the distributions are asymptotically correct when they are estimated on a large number of degrees of freedom. When more than two populations are involved the pooled estimates of the covariances have, usually, a sufficiently large number of degrees of freedom to validate the use of the asymptotic distributions. More exact tests for cases involving small numbers of degrees of freedom are given in the next section.

4. TESTS WITH WILKS' Λ CRITERION

(a) *Analysis of dispersion and the theoretical aspects of the* Λ *criterion*

In the univariate analysis of variance tests of significance reduce to the comparison of two independently distributed mean squares. One of the mean squares is an unbiased estimate of the variance to which a single observation in any particular class is subject and is called the error variance. The other is only so when the null hypothesis which is being tested is correct and may be called the mean square due to deviation from the hypothesis. The test depends only on the individual degrees of freedom of two mean squares.

When each observation consists of p mutually correlated variables there are p total sums of squares and $p(p-1)/2$ total sums of products which can be analysed into various categories. This process which involves the technique of analysing the variances and covariances of multiple correlated variables may be termed the *analysis of dispersion*.* The term

* Prof. R. A. Fisher suggested that this method can be described as either analysis of covariance or analysis of dispersion. Since the term covariance analysis is conventionally used in problems of adjustment for concomitant variation, I have used the term analysis of dispersion to cover a wider variety of problems considered in this article.

dispersion has been used by Prof. Mahalanobis to indicate the scatter of a set of observations as measured by the variances and covariances. Following this terminology the total dispersion may be said to be analysed into dispersion due to various categories.

If we represent the total sum of products by the matrix $S = (S_{ij})$, then the analysis of dispersion consists in analysing each element such as S_{ij}, according to the usual procedure, into various categories with the corresponding distribution of degrees of freedom. The dispersion due to any category supplies the sum of products (denoted for brevity by S.P.) matrix which on division by the degrees of freedom gives the mean product (denoted by M.P.) matrix. The S.P. matrix leading to unbiased estimates of the variances and covariances to which a single set of variables is subject is called the S.P. matrix due to error. This error matrix may be denoted by W with w as its degrees of freedom. In the analysis of dispersion the S.P. matrix due to any other category leads to unbiased estimates of variances and covariances only when the null hypothesis regarding that category is true. This may be called the S.P. matrix due to deviation from the hypothesis. If such a matrix is represented by Q with q as its degrees of freedom, then the problem of testing the null hypothesis consists in comparing the matrices $w^{-1}W$ and $q^{-1}Q$. The simultaneous comparison of the estimates of the variances and covariances appears to be a natural extension of the comparison of variances in the case of a single variate.

The appropriate test criteria for comparison may be obtained by extending the method of discriminant function analysis. A linear compound of the variables is taken and the compounding coefficients are chosen such that the ratio of mean squares due to deviation from hypothesis and due to error for this variable is a maximum. The ratio F^2 which comes out as a root of the determinantal equation

$$\left| Q - \frac{q}{w} F^2 W \right| = 0$$

may be used as the appropriate test criterion. If $|W| \neq 0$, the number of non-zero roots of this equation is equal to the number of variables under consideration or q, the number of degrees of freedom of Q, whichever is smaller. An adequate comparison of $w^{-1}W$ and $q^{-1}Q$ must involve the tests of significance of all the roots. If $F_1^2, F_2^2, \ldots$ represent the various roots, it is easy to verify that

$$\left(1 + \frac{q}{w} F_1^2\right)\left(1 + \frac{q}{w} F_2^2\right) \ldots = \frac{|W+Q|}{|W|}.$$

The ratio $|W|/|W+Q|$ denoted by Λ decreases as the roots increase and a significantly small value of Λ may be taken as providing the significance of one or more of the roots. This is the underlying theory of the Λ criterion arrived at by Wilks (1932) by using the likelihood ratio method and later extended by Bartlett (1934)* for general use in multivariate analysis.

This, however, does not provide a satisfactory test, for when only one or a smaller number of roots than the total indicate real differences, their significance may be obscured by the use of the overall test. Its use can be recommended only in situations where small deviations from the hypothesis can be ignored.

* In a paper read before the Royal Statistical Society in May 1947 (Bartlett 1947), Dr Bartlett suggested a method of factorizing Λ, arising out of a category in the analysis of dispersion, which appeared different from the procedure I have outlined above. In my discussion of Bartlett's paper, I pointed out the difference between his approach in some problems and the general approach by the method of analysis of dispersion, which alone I think can lead to unbiased tests of significance. Such a factorization leads to a valid test in the case $q = 1$ as I have shown elsewhere (Rao, 1946*b*, p. 409, equation 2·14). But this is not true in general.

(b) The distribution of Λ and its practical use

The following notation will be used throughout this and the subsequent sections.

Analysis of dispersion for p variables

Due to	D.F.	S.P. matrix
(1) Deviation from hypothesis	q	Q
(2) Error	$n-q$	W
(3) Total	n	$W+Q$

$$\Lambda = |W|/|W+Q|$$

If the number of variables involved is p, then assuming that the elements of W are distributed independently of those of Q it is easy to derive the tth moment of Λ (Wilks, 1932; Bartlett, 1934) as

$$\prod_{i=0}^{p-1} \frac{\Gamma\{\frac{1}{2}(n-i)\}\,\Gamma\{\frac{1}{2}(n-q-i)+t\}}{\Gamma\{\frac{1}{2}(n-q-i)\}\,\Gamma\{\frac{1}{2}(n-i)+t\}}.$$

The tests based on the exact distributions given by Wilks (1932) and Nair (1939) for some particular cases are reproduced below.

	Nature of test	
	Variance ratic	Degrees of freedom
$q=1$, for any p	$\frac{1-\Lambda}{\Lambda}\frac{n-p}{p}$	p and $(n-p)$
$q=2$, for any p	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\frac{n-p-1}{p}$	$2p$ and $2(n-p-1)$
$p=1$, for any q	$\frac{1-\Lambda}{\Lambda}\frac{n-q}{q}$	q and $(n-q)$
$p=2$, for any q	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\frac{n-q-1}{q}$	$2q$ and $2(n-q-1)$

For other values of p and q, the exact values of the probabilities can be obtained by the use of the χ^2 tables alone as shown below.

It has been shown by Wald & Brookner (1941) that the distribution of the statistic $v = -\log\Lambda$ is of the form

$$2^{\frac{1}{2}pq} C(n)\, e^{-\frac{1}{2}nv} v^{\frac{1}{2}pq-1} \left(\sum_{s=0}^{\infty} \frac{\beta_s v^s}{2^{s+\frac{1}{2}pq}\,\Gamma(s+\frac{1}{2}pq)}\right) dv,$$

where

$$C(n) = \prod_{i=0}^{p-1} \frac{\Gamma\{\frac{1}{2}(n-i)\}}{\Gamma\{\frac{1}{2}(n-q-i)\}},$$

and the β's are the coefficients in the expansion of $\frac{(n/2)^{\frac{1}{2}pq}}{C(n)}$ in powers of $1/n$,

$$\frac{(n/2)^{\frac{1}{2}pq}}{C(n)} = \beta_0 + \frac{\beta_1}{n} + \frac{\beta_2}{n^2} + \ldots.$$

For the purposes of examining Bartlett's approximation and obtaining a quickly convergent series for the tabulation of percentage points, the transformation

$$V = \left(n - \frac{p+q+1}{2}\right) v = mv$$

is made. Changing over to V from v the above distribution becomes

$$2^{\frac{1}{2}pq} D(m)\, m^{-\frac{1}{2}pq} e^{-\frac{1}{2}V} V^{\frac{1}{2}pq-1} \left(\sum_{s=0}^{\infty} \frac{\gamma_s V^s}{m^s 2^{s+\frac{1}{2}pq} \Gamma(s+\frac{1}{2}pq)} \right) dV,$$

where

$$D(m) = \prod_{i=0}^{p-1} \frac{\Gamma\left\{\frac{1}{2}\left(m + \frac{p+q+1}{2} - i\right)\right\}}{\Gamma\left\{\frac{1}{2}\left(m + \frac{p-q+1}{2} - i\right)\right\}}.$$

It is easy to recognize that the γ's depend only on p and q and that they are the coefficients in the expansion of $\frac{(m/2)^{\frac{1}{2}pq}}{D(m)}$ in powers of $1/m$.

The asymptotic expansion of $\frac{(m/2)^{\frac{1}{2}pq}}{D(m)}$ can be calculated from the formulae

$$\left(\frac{m+2}{m}\right)^{\frac{1}{2}pq} \frac{D(m)}{D(m+2)} = \frac{\gamma_0 + \frac{\gamma_1}{m+2} + \dots}{\gamma_0 + \frac{\gamma_1}{m} + \dots},$$

and

$$\frac{D(m)}{D(m+2)} = \prod_{i=0}^{p-1} \frac{m - i + (p-q+1)/2}{m - i + (p+q+1)/2}.$$

Taking the logarithms of the first equality one gets

$$\log\left(\gamma_0 + \frac{\gamma_1}{m+2} + \dots\right) = \frac{pq}{2}\log\left(1 + \frac{2}{m}\right) + \log\left(\gamma_0 + \frac{\gamma_1}{m} + \dots\right) + \sum_i \log\left(1 + \frac{p-q+1-2i}{2m}\right) - \sum_i \log\left(1 + \frac{p+q+1-2i}{2m}\right).$$

Expanding in powers of $(1/m)$ and equating coefficients of like powers on both sides, the first five γ's come out as

$$\gamma_0 = 1,^*$$

$$\gamma_1 = \frac{pq}{2} + \frac{1}{16}\sum_{i=0}^{p-1} \{(p-q+1-2i)^2 - (p+q+1-2i)^2\}$$

$$= 0,$$

$$\gamma_2 = -\frac{pq}{3} + \frac{1}{96}\sum_{i=0}^{p-1} \{(p+q+1-2i)^3 - (p-q+1-2i)^3\}$$

$$= -\frac{pq}{3} + \frac{pq}{48}(p^2+q^2+11)$$

$$= \frac{pq}{48}(p^2+q^2-5),$$

* That the value of $\gamma_0 = 1$ is easily seen by making $m \to \infty$ in the distribution of V in which case it is reduced to the χ^2 form. The comparison of the coefficients, in fact, give the values of the ratios γ_1/γ_0, γ_2/γ_0, ...

$$\gamma_3 = 2\gamma_2 + \frac{pq}{3} + \frac{1}{384}\sum_{i=0}^{p-1}\{(p-q+1-2i)^4 - (p+q+1-2i)^4\}$$

$$= 2\gamma_2 + \frac{pq}{3} - \tfrac{16}{384}pq(p^2+q^2+3)$$

$$= 0,$$

$$\gamma_4 = -4\gamma_2 + \frac{\gamma_2^2}{2} - \tfrac{2}{5}pq + \frac{1}{1280}\sum_{i=0}^{p-1}\{(p+q+1-2i)^5 - (p-q+1-2i)^5\}$$

$$= -4\gamma_2 + \frac{\gamma_2^2}{2} - \tfrac{2}{5}pq + \frac{pq}{1920}\{3p^4+3q^4+110(p^2+q^2)+10p^2q^2+127\}$$

$$= \frac{\gamma_2^2}{2} + \frac{pq}{1920}\{3p^4+3q^4+10p^2q^2-50(p^2+q^2)+150\}.$$

Considering only terms up to the fourth power in $(1/m)$ the distribution of V becomes

$$\left(\frac{2}{m}\right)^{\frac{1}{2}pq} D(m)\, e^{-\frac{1}{2}V} V^{\frac{1}{2}pq-1}\left\{\frac{1}{2^{\frac{1}{2}pq}\Gamma(\frac{1}{2}pq)} + \frac{\gamma_2}{m^2}\frac{V^2}{2^{2+\frac{1}{2}pq}\Gamma(2+\frac{1}{2}pq)} + \frac{\gamma_4}{m^4}\frac{V^4}{2^{4+\frac{1}{2}pq}\Gamma(4+\frac{1}{2}pq)}\right\}.$$

The probability of V exceeding an observed value V_0, then, becomes

$$\left(\frac{2}{m}\right)^{\frac{1}{2}pq} D(m)\left\{P_{pq} + \frac{\gamma_2}{m^2}P_{pq+4} + \frac{\gamma_4}{m^4}P_{pq+8} + \ldots\right\}$$

where

P_{pq+s} = Probability of χ^2 with $pq+s$ degrees of freedom exceeding the value V_0.

One may go a step further and expand $\left(\frac{2}{m}\right)^{\frac{1}{2}pq} D(m)$ in powers of $(1/m)$ in which case the above series becomes

$$P_{pq} + \frac{\gamma_2}{m^2}(P_{pq+4} - P_{pq}) + \frac{1}{m^4}\{\gamma_4(P_{pq+8} - P_{pq}) - \gamma_2^2(P_{pq+4} - P_{pq})\} + \ldots.$$

This form is most convenient for the calculation of the required probability. The quantities γ_2 and γ_4 are simple functions of p and q only. Using χ^2 tables, the γ's and powers of $(1/m)$, the probability of V exceeding V_0 can be calculated to a sufficient degree of accuracy.

Bartlett (1938) suggested the use of

$$V = -m\log_e\Lambda = -\left(n - \frac{p+q+1}{2}\right)\log_e\Lambda$$

as χ^2 with pq degrees of freedom. This corresponds to using the first term of the series. Since the second term is $o(1/m^2)$, its contribution is very small even for moderately large m so that, in many practical problems, Bartlett's approximation can be safely used. For small values of m one may use the second and the third terms depending on the accuracy needed in any problem.

(c) *Test of differences in mean values for several populations*

Let $\pi_1, \ldots, \pi_k$ be k populations from which samples of sizes $N_1, \ldots, N_k$ for p correlated variables are available. The dispersion has to be analysed into 'between' and 'within' populations. The S.P. matrix 'within' populations (the error) has $N_1 + \ldots + N_k - k$ degrees of freedom and that 'between' populations has $(k-1)$ degrees of freedom. If these are

represented by W and Q then the statistic to be used for testing the differences in mean values is

$$V = -m \log_e \Lambda,$$

where

$$\Lambda = |W| / |W+Q|,$$

$$m = n - \frac{p+q+1}{2},$$

$$n = N_1 + \ldots + N_k - 1,$$

$$q = k-1.$$

The exact probability of V exceeding the observed value can be calculated as explained in § 4 (*b*).

Example 3. Table 4 gives the analysis of dispersion for the three characters, head length (x_1), height (x_2) and weight (x_3) measured on 140 schoolboys, of almost the same age, belonging to six different schools in an Indian city.

Table 4. *Analysis of dispersion*

Dispersion	D.F.	S.P. matrix					
		x_1^2	x_2^2	x_3^2	x_1x_2	x_1x_3	x_2x_3
'Between' schools	5	(Q_{ij}) 752·0	151·3	1612·7	214·2	521·3	401·2
'Within' schools	134	(W_{ij}) 12809·3	1499·6	21009·6	1003·7	2671·2	4123·6
Total	139	(S_{ij}) 13561·3	1650·9	22622·3	1217·9	3192·5	4524·8

$$\Lambda = \frac{|W|}{|S|} = \frac{\begin{vmatrix} 12809{\cdot}3 & 214{\cdot}2 & 521{\cdot}3 \\ 214{\cdot}2 & 1499{\cdot}6 & 401{\cdot}2 \\ 521{\cdot}3 & 401{\cdot}2 & 21009{\cdot}6 \end{vmatrix}}{\begin{vmatrix} 13561{\cdot}3 & 1217{\cdot}9 & 3192{\cdot}5 \\ 1217{\cdot}9 & 1650{\cdot}9 & 4524{\cdot}8 \\ 3192{\cdot}5 & 4524{\cdot}8 & 22622{\cdot}3 \end{vmatrix}}$$

$$= \frac{10^{12}(0{\cdot}176005)}{10^{12}(0{\cdot}213628)},$$

$$-\log_e \Lambda = 0{\cdot}193724,$$

$$m = 139 - \tfrac{1}{2}(5+3+1) = 134{\cdot}5,$$

$$V = -m \log_e \Lambda = (134{\cdot}5)(0{\cdot}193724)$$

$$= 26{\cdot}0559.$$

Using V as χ^2 with $pq = 15$ degrees of freedom the first approximation comes out as

$$P_{15} = 0{\cdot}0375.$$

The second term is

$$\frac{\gamma_2}{m^2}(P_{19} - P_{15}),$$

where

$$\gamma_2 = \frac{29 \times 15}{48} \quad \text{and} \quad \frac{\gamma_2}{m^2} = \frac{29 \times 15}{48(134{\cdot}5)^2} = 0{\cdot}00050096,$$

$$\frac{\gamma_2}{m^2}(P_{19} - P_{15}) = 0{\cdot}00050096(0{\cdot}1285 - 0{\cdot}0375)$$

$$= 0{\cdot}00004574.$$

This correction to the first approximation affects only the fourth decimal place so that correction is hardly necessary. The observed value of V is significant at the 5 % level showing thereby that boys of various schools differ in physique. This appears to be generally true since boys belonging to different social strata attend different schools.

(*d*) *Internal analysis of a set of variates*

Let $x_1, \ldots, x_s, x_{s+1}, \ldots, x_{s+p}$ be $(s+p)$ correlated variables for which samples of sizes $N_1, \ldots, N_k$ are available from k populations. If the differences in mean values of these $(s+p)$ variables are to be tested for significance, then the method given in § 4 (*c*) can be used. An important problem that arises in biometry is to test whether the variables, say $x_{s+1}, \ldots, x_{s+p}$ bring out further differences in populations when the differences due to $x_1, \ldots, x_s$ are removed.

It is apparent in problems of this nature that some of the variables in the set $x_1, \ldots, x_s$ might be in the nature of concomitant variates which have been observed in association with the dependent variables or which might have been chosen to have some specified values. An illustration of such an analysis is given in my discussion on Bartlett's paper (Bartlett, 1947). In that problem there were three dependent variables g, h and i corresponding to linear, parabolic and cubic terms of growth curves of pig weights and a concomitant variable w giving the initial weight of pigs. It was desired to test whether the variables h and i bring out further differences in food treatments when the differences due to g and w are eliminated. The problem is identical with that posed above with g, w forming the first set and h, i the second set of variables.

There is a third set of problems in which it is desired to test whether the differences in k groups characterized by $(s+p)$ characters can be explained by variations in s *assigned* linear functions of these variables. If $y_1, \ldots, y_{s+p}$ are the $(s+p)$ variables and

$$L_1 = m_{1,1}y_1 + \ldots + m_{1,\,p+s}y_{p+s},$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$L_s = m_{s,1}y_1 + \ldots + m_{s,\,p+s}y_{p+s},$$

are the assigned linear functions, then one can replace the $(s+p)$ variables $y_1, \ldots, y_{s+p}$ by $x_1, \ldots, x_{s+p}$ defined by

$$x_1 = L_1, \quad \ldots, \quad x_s = L_s,$$
$$x_{s+1} = m_{s+1,1}y_1 + \ldots + m_{s+1,\,s+p}y_{s+p},$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$x_{s+p} = m_{s+p,1}y_1 + \ldots + m_{s+p,\,s+p}y_{s+p},$$

where the coefficients in $x_{s+1}, \ldots, x_{s+p}$ are chosen arbitrarily subject to the condition that the determinant $|\, m_{ij} \,|$, $i, j = 1, 2, \ldots, (s+p)$ is not zero. This latter condition ensures that the transformation from the y's to the x's leads to one-to-one correspondence. Once again, the problem is reduced to that of considering the differences in $x_{s+1}, \ldots, x_{s+p}$ when those due to $x_1, \ldots, x_s$ are removed. The proposed test is independent of the compounding coefficients used to define the set $x_{s+1}, \ldots, x_{s+p}$ so that, in any practical problem, they may be conveniently or conventionally chosen.

In all these cases, the problem is one of analysing the dispersion of the variables $x_{s+1}, \ldots, x_{s+p}$ when the dispersion due to $x_1, \ldots, x_s$ is removed. This can be done by following the covariance technique suitable for p dependent variables and s independent variables (Wishart, 1936, 1939; Rao, 1946*a*).

Let $(S_{ij}) = (Q_{ij}) + (W_{ij}) \quad (i, j = 1, 2, \ldots, (s+p))$

be the analysis of dispersion for all the $(s+p)$ variates due to deviation from hypothesis and error with the corresponding distribution of degrees of freedom as

$$n' = q + (n' - q).$$

The S.P. matrix due to error for the variables $x_1, \ldots, x_s$ to be eliminated is

$$\begin{pmatrix} W_{11} & \cdots & W_{1s} \\ \cdots & \cdots & \cdots \\ W_{s1} & \cdots & W_{ss} \end{pmatrix}$$

and its inverse is represented by $\begin{pmatrix} W^{11} & \cdots & W^{1s} \\ \cdots & \cdots & \cdots \\ W^{s1} & \cdots & W^{ss} \end{pmatrix}$

The S.P. matrix due to error for $x_{s+1}, \ldots, x_{s+p}$ when corrected for $x_1, \ldots, x_s$, is given by $W(s+1, \ldots, s+p \mid 1, \ldots, s)$ or simply $W(p \mid s)$ where

$$W(p \mid s) = \begin{pmatrix} W_{s+1,s+1} & \cdots & W_{s+1,s+p} \\ \cdots & \cdots & \cdots \\ W_{s+p,s+1} & \cdots & W_{s+p,s+p} \end{pmatrix} - \begin{pmatrix} W_{1,s+1} & \cdots & W_{s,s+1} \\ \cdots & \cdots & \cdots \\ W_{1,s+p} & \cdots & W_{s,s+p} \end{pmatrix} \begin{pmatrix} W^{11} & \cdots & W^{1s} \\ \cdots & \cdots & \cdots \\ W^{s1} & \cdots & W^{ss} \end{pmatrix} \begin{pmatrix} W_{1,s+1} & \cdots & W_{1,s+p} \\ \cdots & \cdots & \cdots \\ W_{s,s+1} & \cdots & W_{s,s+p} \end{pmatrix}.$$

This form which involves the evaluation of a triple product of matrices appears to be most convenient for computation as illustrated in the next section. Replacing W by S one has the formula for computing the S.P. matrix due to 'deviation from hypothesis + error' for $x_{s+1}, \ldots, x_{s+p}$ when corrected for $x_1, \ldots, x_s$. If this is represented by $S(p \mid s)$ then the required criterion is

$$\Lambda = \frac{|W(p \mid s)|}{|S(p \mid s)|}.$$

The degrees of freedom for $W(p \mid s)$ are $(n' - q - s)$ and for $S(p \mid s)$ are $(n' - s)$ so that in standard notation the parameters associated with Λ are

$$n = n' - s, \quad p = p, \quad q = q.$$

The test car be carried out as discussed in § 4 (*b*).

(*c*) *Barnard's problem of secular variations in skull characters*

The problem of measuring secular variations in skull characters considered by Barnard (1935) is of immense importance to the anthropologists. It is, however, of interest to examine the methods employed by her in the light of latest developments in multivariate analysis. The two problems involved in her study are

(i) the selection of a smaller number, out of seven skull characters, which give significant information, so far as is possible, as to changes taking place with time in four series of Egyptian skulls, and

(ii) the determination of an expression, linear in measurements, which characterizes most effectively an individual skull with respect to the progressive secular changes.

To answer problem (i) Barnard first chose Basi-alveolar Length and Nasal Height as two basic characters which independently of each other, show significant variation in the four

series. To choose further characters she considered the problem of testing the significance of the linear regression of the mean values of an added character with time (corresponding to the four series) when that part of the regression due to the two basic characters is removed. This meant the choice of characters with special reference to the average linear rate of change of the individual means with time. If the choice of characters is to be with reference to the complete nature of changes taking place with time, then what is needed is an internal analysis of the characters to decide whether the configuration of the four series as determined by several characters is the same as that indicated by a smaller number. Barnard's method should, of course, be preferred if the regressions were known to be linear. This can, however, be tested from the data.

Taking the four measurements

x_1 Basi-alveolar Length,
x_2 Nasal Height,
x_3 Maximum Breadth,
x_4 Basi-bregmatic Height,

the relevant data are summarized in Tables 5 and 6 which give the means for the four series and the analysis of dispersion.

Table 5. *Means for the four series*

Character	Series			
	I $N_1 = 91$	II $N_2 = 162$	III $N_3 = 70$	IV $N_4 = 75$
x_1	133·582418	134·265432	134·371429	135·306667
x_2	98·307692	95·462963	95·857143	95·040000
x_3	50·835165	51·148148	50·100000	52·093333
x_4	133·000000	134·882716	133·642857	131·466667

Table 6. *Analysis of dispersion*

	Dispersion		
	'Between' (3 D.F.)	'Within' (394 D.F.)	Total (397 D.F.)
x_1^2	123·180628	9661·997470	9785·178098
x_2^2	486·345863	9073·115027	9559·460890
x_3^2	100·411505	3938·320351	4088·731856
x_4^2	640·733891	8741·508829	9382·242720
x_1x_2	−231·375635	445·573301	214·197666
x_1x_3	87·305348	1130·623900	1217·929248
x_1x_4	−128·763994	2148·584210	2019·820216
x_2x_3	−107·505618	1239·221990	1131·716372
x_2x_4	125·313318	2255·812722	2381·126040
x_3x_4	−137·580764	1271·054662	1133·473898

Example 4. Do the characters x_3 and x_4 show significant variation in the four series independently of the variation due to the characters x_1 and x_2?

The method developed in § 4 (*d*) is directly useful in this problem. The S.P. matrix 'within' for the basic characters x_1 and x_2 is

$$\begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} = \begin{pmatrix} 9661{\cdot}997470 & 445{\cdot}573301 \\ 445{\cdot}573301 & 9073{\cdot}115027 \end{pmatrix}.$$

Its inverse is

$$\begin{pmatrix} W^{11} & W^{12} \\ W^{21} & W^{22} \end{pmatrix} = 10^{-4}\begin{pmatrix} 1{\cdot}037332 & -0{\cdot}050942 \\ -0{\cdot}050942 & 1{\cdot}104659 \end{pmatrix}.$$

The 'within' S.P. matrix for x_3, x_4 due to x_1, x_2 is given by the triple product

$$\begin{aligned} &\begin{pmatrix} W_{13} & W_{23} \\ W_{14} & W_{24} \end{pmatrix}\begin{pmatrix} W^{11} & W^{12} \\ W^{21} & W^{22} \end{pmatrix}\begin{pmatrix} W_{13} & W_{14} \\ W_{23} & W_{24} \end{pmatrix} \\ &\quad = \begin{pmatrix} 1130{\cdot}623900 & 1239{\cdot}221990 \\ 2148{\cdot}584210 & 2255{\cdot}812722 \end{pmatrix}\begin{pmatrix} W^{11} & W^{12} \\ W^{21} & W^{22} \end{pmatrix}\begin{pmatrix} W_{13} & W_{14} \\ W_{23} & W_{24} \end{pmatrix} \\ &\quad = 10^{-4}\begin{pmatrix} 1109{\cdot}703904 & 1311{\cdot}321492 \\ 2113{\cdot}879535 & 2382{\cdot}450625 \end{pmatrix}\begin{pmatrix} W_{13} & W_{14} \\ W_{23} & W_{24} \end{pmatrix} \\ &\quad = \begin{pmatrix} 287{\cdot}967620 & 534{\cdot}238796 \\ 534{\cdot}238796 & 991{\cdot}621041 \end{pmatrix}. \end{aligned}$$

The 'within' S.P. matrix for x_3 and x_4 after correcting for x_1 and x_2 is

$$\begin{aligned} &\begin{pmatrix} W_{33} & W_{34} \\ W_{34} & W_{44} \end{pmatrix} - \begin{pmatrix} W_{13} & W_{23} \\ W_{14} & W_{24} \end{pmatrix}\begin{pmatrix} W^{11} & W^{12} \\ W^{21} & W^{22} \end{pmatrix}\begin{pmatrix} W_{13} & W_{14} \\ W_{23} & W_{24} \end{pmatrix} \\ &\quad = \begin{pmatrix} 3938{\cdot}320351 & 1271{\cdot}054662 \\ 1271{\cdot}054662 & 8741{\cdot}508829 \end{pmatrix} - \begin{pmatrix} 287{\cdot}967620 & 534{\cdot}238796 \\ 534{\cdot}238796 & 991{\cdot}621041 \end{pmatrix} \\ &\quad = \begin{pmatrix} 3650{\cdot}353731 & 736{\cdot}815866 \\ 736{\cdot}815866 & 7749{\cdot}887788 \end{pmatrix} = W(2\,|\,2). \end{aligned}$$

This has $394-2=392$ degrees of freedom. Similarly $S(2\,|\,2)$ with $397-2=395$ degrees of freedom is

$$\begin{pmatrix} 3809{\cdot}335190 & 611{\cdot}798381 \\ 611{\cdot}798381 & 8393{\cdot}755848 \end{pmatrix},$$

$$\Lambda = \frac{|W(2\,|\,2)|}{|S(2\,|\,2)|} = \frac{0{\cdot}27746934}{0{\cdot}31600332} = 0{\cdot}878058,$$

$$V = -m\log_e\Lambda, \quad m = n - \frac{p+q+1}{2} = 395 - \frac{2+3+1}{2} = 392,$$

$$V = -392\log_e(0{\cdot}878058) = 51{\cdot}39.$$

This value of V on $pq=6$ degrees of freedom is highly significant so that x_3 and x_4 may be considered as discriminating the series independently of x_1 and x_2.

Example 5. Taking the relative times between the series in the proportion 2 : 1 : 2, can the variation of the characters be accounted for by the linear regression of individual characters with time?

In order to obtain the regression with time the values of t, the time variable, may be taken as -5, -1, $+1$ and $+5$ for the individuals of the first, second, third and fourth series respectively. The calculation of individual regressions involves the quantities

$$\Sigma(t-\bar{t})^2 = 4307{\cdot}66832,$$

$$\Sigma x_1(t-\bar{t}) = \quad 718{\cdot}76286, \quad \Sigma x_3(t-\bar{t}) = -410{\cdot}10194,$$

$$\Sigma x_2(t-\bar{t}) = -1407{\cdot}26075, \quad \Sigma x_4(t-\bar{t}) = -733{\cdot}42758.$$

The matrix R with 1 degree of freedom giving the squares and products due to regression is given in Table 7.

Table 7. *Matrix R with* 1 *degree of freedom*

	x_1	x_2	x_3	x_4
x_1	119·930358	−234·810812	68·428235	−122·377258
x_2	−234·810812	459·734449	−133·975163	−149·601596
x_3	68·428235	−133·975163	39·042852	− 69·824358
x_4	−122·377258	−149·601596	− 69·824358	124·874099

In the above table

$$R_{11} = [\Sigma x_1(t-\bar{t})]^2/\Sigma(t-\bar{t})^2, \quad R_{12} = [\Sigma x_1(t-\bar{t})]\,[\Sigma x_2(t-\bar{t})]/\Sigma(t-\bar{t})^2,$$

and so on. With these results one may analyse the dispersion of which a typical product (x_1x_2) is chosen below for illustration.

Table 8. *Analysis of dispersion*

Due to	D.F.	S.P. matrix (x_1x_2)
Regression	1	−234·810812 (R_{ij})
*Deviation from regression	2	3·435177 (Q_{ij})
Total ('between' series)	3	−231·375635 $(R_{ij}+Q_{ij})$
'Within' series	394	445·573301 (W_{ij})
Total	397	214·197666 (S_{ij})
Deviation from regression + 'within'	396	449·008478 $(Q_{ij}+W_{ij})$

* This quantity is obtained by subtraction. The complete matrix $(Q_{ij}+W_{ij})$ obtained by the above method is given in table 9.

Table 9. *Matrix* $(Q_{ij}+W_{ij})$ *with* 396 *degrees of freedom*

x_1	x_2	x_3	x_4
9665·247740	449·008478	1149·501013	2142·197474
449·008478	9099·726441	1265·691535	2231·524444
1149·501013	1265·691535	4049·689004	1203·298256
2142·197474	2231·524444	1203·298256	9257·368621

To test the hypothesis that the regressions are linear one has to compare W and $Q+W$.

$$\Lambda = \frac{|W|}{|Q+W|} = \frac{0{\cdot}24269054\times10^{12}}{0{\cdot}26873816\times10^{12}} = 0{\cdot}90307436,$$

$$V = -\left\{396-\frac{2+4+1}{2}\right\}\log_e(0{\cdot}90307436)$$

$$= 40{\cdot}02.$$

The χ^2 approximation has $p \times q = 2 \times 4 = 8$ degrees of freedom, since Q has 2 degrees of freedom and there are four variables. The result is significant so that the regressions cannot be considered linear.

This test can be extended to examine whether a parabolic regression with time can explain the differences in mean values. The matrix Q giving the deviation from regression has then 1 degree of freedom and R due to regression 2.

To determine the coefficients of a linear compound which characterizes most effectively the secular changes in progress, Barnard maximized the ratio of the square of unweighted regression of the compound with time. It is doubtful, as Bartlett (1947) points out, whether such a linear compound can be used to specify an individual skull most effectively with respect to progressive changes, since linear regression with time does not adequately explain all the differences in the four series.

The variance ratios obtained in this paper in the case of two samples can also be derived from a general regression analysis by considering certain pseudo-variates, which have constant values for members of the same sample, as dependent variables and the observed values as independent variables (Bartlett, 1939; Fisher, 1940; Brown, 1947). Such an approach does not seem to be possible in the case of a single sample. On the other hand, the statistics defined in (2·1) and (2·2) can be used in any situation where there are a number of linear hypotheses to be tested with the use of the estimated deviations and their independently estimated dispersion matrix. The distribution of the statistic in (2·2) has been derived by the author (Rao, 1946*b*) under these general conditions.

I wish to thank Dr Wishart for his helpful criticism during the preparation of this paper.

REFERENCES

Barnard, M. M. (1935). The secular variation of skull characters in four series of Egyptian skulls. *Ann. Eugen., Lond.*, **7**, 89.

Bartlett, M. S. (1934). The vector representation of a sample. *Proc. Camb. Phil. Soc.* **30**, 327.

Bartlett, M. S. (1938). Further aspects of the theory of multiple regression. *Proc. Camb. Phil. Soc.* **34**, 33.

Bartlett, M. S. (1939). The standard errors of discriminant function coefficients. *J. Roy. Statist. Soc. Suppl.* **6**, 169.

Bartlett, M. S. (1947). Multivariate analysis. *J. Roy. Statist. Soc. Suppl.* **9**, 76.

Bose, R. C. & Roy, S. N. (1938). The distribution of the studentized D^2-statistic. *Sankhyā*, **4**, 19.

Brown, G. W. (1947). Discriminant functions. *Ann. Math. Statist.* **18**, 514.

Fairfield Smith (1936). A discriminant function for plant selection. *Ann. Eugen., Lond.*, **7**, 240.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen., Lond.*, **7**, 179.

Fisher, R. A. (1938). The statistical utilization of multiple measurements. *Ann. Eugen., Lond.*, **8**, 376.

Fisher, R. A. (1939). The sampling distribution of some statistics obtained from nonlinear regression. *Ann. Eugen., Lond.*, **9**, 238.

Fisher, R. A. (1940). The precision of the discriminant function. *Ann. Eugen., Lond.*, **10**, 422.

Hotelling, H. (1931). The generalization of Student's ratio. *Ann. Math. Statist.* **2**, 360.

Hotelling, H. (1935). The relation between two sets of variates. *Biometrika*, **28**, 321.

Hotelling, H. (1947). A generalized T measure of multivariate dispersion. *Ann. Math. Statist.* **18**, 298 (Abstract 2).

Hsu, P. (1939*a*). On the generalized analysis of variance. *Biometrika*, **31**, 221.

Hsu, P. (1939*b*). On the distribution of the roots of certain determinantal equations. *Ann. Eugen., Lond.*, **9**, 250.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proc. Nat. Inst. Sci.* (India), **12**, 49.

MARTIN, E. A. (1935). A study of the Egyptian series of mandibles, with special reference to mathematical methods of sexing. *Biometrika*, **28**, 149.

NAIR, U. S. (1939). The application of moment functions in the study of distribution laws in statistics. *Biometrika*, **30**, 274.

NEYMAN, J. & PEARSON, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, **20A**, 175.

NEYMAN, J. & PEARSON, E. S. (1931). On the problem of k samples. *Bull. int. Acad. Cracovie*, A, p. 460.

PEARSON, E. S. & NEYMAN, J. (1930). On the problem of two samples. *Bull. int. Acad. Cracovie*, A, p. 73.

PLACKETT, R. L. (1947). An exact test for the equality of variances. *Biometrika*, **34**, 311.

RAO, C. R. (1945). Generalization of Markoff's theorem and tests of linear hypothesis. *Sankhyā*, **7**, 9.

RAO, C. R. (1946*a*). On the linear combination of observations and the general theory of least squares. *Sankhyā*, **7**, 237.

RAO, C. R. (1946*b*). Tests with discriminant functions in multivariate analysis. *Sankhyā*, **7**, 407.

RAO, C. R. (1947). The problem of classification and distance between two populations. Letter to *Nature*, **159**, 30.

ROY, S. N. (1939). p-statistics or some generalizations in analysis of variance appropriate to multivariate problems. *Sankhyā*, **4**, 381.

WALD, A. & BROOKNER, R. J. (1941). On the distribution of Wilks' statistic for testing independence of several groups of variables. *Ann. Math. Statist.* **12**, 137.

WELCH, B. L. (1939). Note on discriminant functions. *Biometrika*, **31**, 218.

WILKS, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, **24**, 471.

WILKS, S. S. (1935). On the independence of k sets of normally distributed statistical variables. *Econometrica*, **3**, 309.

WISHART, J. (1928). The generalized product moment distribution in samples from a normal multivariate population. *Biometrika*, **20A**, 32.

WISHART, J. (1936). Tests of significance in analysis of covariance. *J. Roy. Statist. Soc. Suppl.* **3**, 79.

WISHART, J. (1939). Statistical treatment of animal experiments. *J. Roy. Statist. Soc. Suppl.* **6**, 1.

ALTERNATIVE SYSTEMS IN THE ANALYSIS OF VARIANCE

By N. L. JOHNSON

1. It is the purpose of this paper to compare the fundamental theoretical set-ups implied by certain well-known systems of approach to the analysis of variance. Differences in interpretation in the different systems are discussed, and attention is drawn to some particularly simple results in the theory associated with one of the systems. No attempt is made to place the systems in any order of general preference. It is the author's opinion that each has its own sphere of application, while consideration of problems from the viewpoints of more than one of the systems will often prove enlightening.

2. The power functions of tests used in the analysis of variance have been considered by Hsu (1941), and by Tang (1938) who has given tables by means of which the power may be evaluated numerically in certain cases.

In the particular case of testing for differences between k groups, the theoretical set-up used by these authors is

$$x_{ij} = B_i + z_{ij} \quad (i = 1, \ldots, k; j = 1, \ldots, n_i), \tag{1}$$

where x_{ij} corresponds to the jth observation in the ith group. B_i is a constant representing the expected value in the ith group and the z_{ij}'s are independent random variables, each with zero expected value and standard deviation σ.

The appropriate analysis of variance is:

Source	Sum of squares	Degrees of freedom	Mean square
Between groups	$\sum_{i=1}^{k} n_i(\bar{x}_{i.}-\bar{x}_{..})^2$	$k-1$	$\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{x}_{i.}-\bar{x}_{..})^2$
Within groups	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i.})^2$	$N-k$	$\frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i.})^2$

where $N = \sum_{i=1}^{k} n_i$ is the total number of observations; $\bar{x}_{i.} = n_i^{-1}\sum_{j=1}^{n_i} x_{ij}$ is the mean observed value in the ith group; $\bar{x}_{..} = N^{-1}\sum_{i=1}^{k} n_i\bar{x}_{i.}$ is the mean of all the observed values. The expected value of the between groups mean square is $\sigma^2 + \frac{1}{k-1}\sum_{i=1}^{k} n_i(B_i-\bar{B})^2$. The expected value of the within groups mean square is σ^2. $\left(\bar{B} = N^{-1}\sum_{i=1}^{k} n_i B_i.\right)$ The ratio

(between groups mean square)/(within groups mean square)

is used to test whether there is any difference between the B_i's. If there is no such difference the expected values of the numerator and denominator of the above ratio are each equal to σ^2; otherwise the expected value of the numerator is greater than σ^2.

If it be assumed that each of the z_{ij}'s is normally distributed, the ratio of the mean squares will be referred to the F distribution with degrees of freedom $\nu_1 = k-1$, $\nu_2 = N-k$. Large observed values of the ratio are regarded as significant. A suitable upper significance limit of the appropriate F distribution may be used as a formal critical limit for the mean square ratio. Tang showed that, for such a test with alternative hypotheses specified by (1) above, the power (i.e. the chance of establishing significance when the B_i's are not all equal) depends

on k, N and, say, $\theta = (k-1)^{-1}\sum_{i=1}^{k} n_i(B_i-\bar{B})^2/\sigma^2$ only. The power function is somewhat complicated in its mathematical expression, but Tang's tables make it possible to determine the chance that a given ratio θ would be established as significant at either the 5 or 1 % level.

3. The alternative approach described below involves a modification in the theoretical set-up and leads to a very simple form of power function in a particular, but common, case. Although the modification may not always be justifiable, it will often provide a more accurate model than set-up (1), apart from the greater simplicity in the resulting analysis.

This alternative form of theoretical set-up is constructed by replacing (1) by

$$x_{ij} = A + z'_i + z_{ij}, \tag{2}$$

where A is a constant and the z'_i's are independent random variables, each with expected value zero and standard deviation σ'. The z'_i's and the z_{ij}'s are also mutually independent. In (2) the constants B_i of (1) are replaced by the random variables $A+z'_i$, and the hypothesis $B_1 = B_2 = \ldots = B_k$ is replaced by the hypothesis $\sigma' = 0$.

It is evident that (2) is a suitable set-up if it is possible to regard the groups as being chosen at random from a large assemblage of groups. If the groups are fixed, (1) will be preferable. Sub-sampling by batches from a randomly selected sample of batches is a typical example where (2) is suitable; comparison of a number of distinct treatments of a material is a typical example where (1) is preferable. Set-ups (1) and (2) represent extreme cases. In general it is likely that an intermediate set-up of form

$$x_{ij} = B_i + z'_i + z_{ij}$$

would be most appropriate. Daniels (1939) and Eisenhart (1947) have discussed, in some detail, the factors affecting the relevance of set-ups (1) and (2) in any particular problem.

However, the same analysis of variance is suitable in both extreme cases and so is likely to be suitable for any intermediate case. In fact, under the assumptions summarized in (2):

The expected value of the between groups mean square is $\sigma^2 + \sigma'^2\left(N^2 - \sum_{i=1}^{k} n_i^2\right)\Big/N(k-1)$.

The expected value of the within groups mean square is σ^2. The ratio

(between groups mean square)/(within groups mean square)

is again suitable for testing the hypothesis $\sigma' = 0$. If it be assumed that the z_{ij}'s are normally distributed, the F significance limits may be used as described in § 2.

4. If it now be assumed that the z'_i's, as well as the z_{ij}'s, are normally distributed, it is possible to obtain the power function in the case $n_1 = n_2 = \ldots = n_k = n$, say, in a particularly simple form. From (2) it follows that

$$\bar{x}_{i.} = A + z'_i + \bar{z}_{i.} \quad \text{where} \quad \bar{z}_{i.} = n^{-1}\sum_{j=1}^{n} z_{ij}$$

$$= A + u_i \qquad \text{where} \quad u_i = z'_i + \bar{z}_{i.}.$$

The u_i's are independent normal variables each with expected value zero and standard deviation $\sqrt{(n^{-1}\sigma^2 + \sigma'^2)}$. The between groups sum of squares is

$$\sum_{i=1}^{k} n(\bar{x}_{i.} - \bar{x}_{..})^2 = \sum_{i=1}^{k} n(u_i - \bar{u})^2 \quad \text{where} \quad \bar{u} = k^{-1}\sum_{i=1}^{k} u_i,$$

and hence is distributed as $\chi^2(\sigma^2 + n\sigma'^2)$ with degrees of freedom $\nu = k-1$. The within groups sum of squares is

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i.})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(z_{ij} - \bar{z}_{i.})^2,$$

and is distributed independently as $\chi^2\sigma^2$ with $\nu_2 = k(n-1)$ degrees of freedom. Hence the mean square ratio used in the analysis of variance test is distributed as

$$F \times \left(\frac{\sigma^2 + n\sigma'^2}{\sigma^2}\right),$$

with $\nu_1 = k-1$, $\nu_2 = k(n-1)$.

If F_α, the upper 100α % point of the F distribution, be used as a formal critical limit, the probability of rejection of the hypothesis $\sigma' = 0$ is

$$\Pr.\left\{F\left(\frac{\sigma^2+n\sigma'^2}{\sigma^2}\right) > F_\alpha\right\} = \Pr.\left\{F > F_\alpha\left(1+\frac{n\sigma'^2}{\sigma^2}\right)^{-1}\right\}. \tag{3}$$

This is the power of the test with respect to the alternative hypothesis specified by (2) and a particular (non-zero) value of σ'. When $\sigma' = 0$ the probability of rejection is α. As σ' increases $F_\alpha(1+n\sigma'^2/\sigma^2)^{-1}$ decreases and the power increases. The expression (3), considered as a function of σ', is the power function of the analysis of variance test in this case.

It will be noted that the power function is a function of $n\sigma'^2/\sigma^2$. This is analogous to the fact that with set-up (1) the power function is a function of $(n/(k-1))\sum_{i=1}^{k}(B_i-\bar{B})^2/\sigma^2$. σ'^2 and $\sum_{i=1}^{k}(B_i-\bar{B})^2/(k-1)$, indeed, fulfil similar roles in the two systems.

5. The calculation of the power function from (3) is straightforward. Since

$$p(F) = \frac{(k-1)^{\frac{1}{2}(k-1)}[k(n-1)]^{\frac{1}{2}k(n-1)}}{B(\frac{1}{2}[k-1], \frac{1}{2}k[n-1])} \cdot \frac{F^{\frac{1}{2}(k-3)}}{[k(n-1)+(k-1)F]^{\frac{1}{2}(nk-1)}},$$

it follows that Pr. $\{F > F_\alpha(1+n\sigma'^2/\sigma^2)^{-1}\}$ can be expressed as the incomplete beta function ratio

$$I_\phi(\tfrac{1}{2}k[n-1], \tfrac{1}{2}[k-1]),$$

where $$\phi = k(n-1)/\{k(n-1)+(k-1)F_\alpha(1+n\sigma'^2/\sigma^2)^{-1}\}.$$

When the number of groups is odd and not large it is possible to evaluate (3) as a simple explicit function $\beta(\sigma'/\sigma)$ of σ'/σ. For example if $k = 3$,

$$p(F) = K_n[\tfrac{3}{2}(n-1)+F]^{-\frac{1}{2}(3n-1)},$$

where $$K_n = [\tfrac{3}{2}(n-1)]^{\frac{1}{2}(3n-5)}.$$

Hence
$$\beta(\sigma'/\sigma) = \int_{F_\alpha(1+n\sigma'^2/\sigma^2)^{-1}}^{\infty} p(F)\,dF$$
$$= K'_n[\tfrac{3}{2}(n-1)+F_\alpha(1+n\sigma'^2/\sigma^2)^{-1}]^{-\frac{1}{2}(n-1)}, \tag{4}$$

where $$K'_n = K_n/\tfrac{3}{2}(n-1) = [\tfrac{3}{2}(n-1)]^{\frac{1}{2}(3n-7)}.$$

Since
$$\int_{F_\alpha}^{\infty} p(F)\,dF = \alpha,$$
$$K'_n[\tfrac{3}{2}(n-1)+F_\alpha]^{-\frac{1}{2}(n-1)} = \alpha,$$

and so
$$\beta(\sigma'/\sigma) = \alpha\left[\frac{\frac{3}{2}(n-1)+F_\alpha}{\frac{3}{2}(n-1)+F_\alpha(1+n\sigma'^2/\sigma^2)^{-1}}\right]^{\frac{1}{2}(n-1)} \tag{4 bis}$$

6. In many cases the hypothesis $B_1 = B_2 = \ldots = B_k$, or the corresponding hypothesis in set-up (2), $\sigma' = 0$, is unduly stringent. That is to say, a test procedure is required which will lead to acceptance in a high proportion of cases provided there is not *too much* difference between the groups. It is natural to use as measure of the amount of difference between the

groups a parameter on which the power function depends. The parameter to be used would therefore be $\sum_{i=1}^{k} n_i(B_i-\bar{B})^2/(k-1)$ in conjunction with set-up (1), or σ' with set-up (2). The critical limit for the mean square ratio should then be chosen so that the chance of rejection is less than α_1, say, for any value of the parameter less than some specified amount.

It will also be desirable that rejection shall take place in a high proportion of cases—at least $1-\alpha_2$, say—when the groups differ by more than a certain amount. That is, the test should have at least a certain minimum sensitivity.

The requirements described above may be summed up formally as follows:

(*a*) The power function of the test must be less than α_1 for all values of the difference parameter less than some known amount.

(*b*) The power function must be greater than $1-\alpha_2$ for all values of the difference parameter greater than a second known amount.

This type of problem occurs in the theory of statistical quality control, the limiting values of the parameter corresponding to 'Producer's Tolerance' or 'Process Average' and 'Consumer's Tolerance' or 'Lot Tolerance' respectively.

Evidently, conditions (*a*) and (*b*) cannot both be satisfied unless the data available are sufficiently numerous. The amount of data required can be found from a study of the appropriate power functions. Tang's tables are useful in this connexion when set-up (1) is suitable. The analysis is much simpler for set-up (2) in the case we have considered in § 3 *et seq.* In the next section certain interesting results will be derived for this particular situation.

7. It will be convenient to define the difference between the groups by the parameter σ'/σ, instead of simply σ'. It may be noted that σ'/σ is the ratio of a parameter representing between-groups variability to one representing within-groups variability, and so can be regarded as a measure of *relative* between-groups variability. We shall require the analysis of variance test to be such that

(*a*) the probability of rejection is less than α_1 if $\sigma'/\sigma < \lambda_1$;

(*b*) the probability of rejection is greater than $1-\alpha_2$ if $\sigma'/\sigma > \lambda_2$.

In the limiting case when (*a*) is just satisfied, it is clear that the critical limit F_0 must be such that

$$\Pr.\{F(1+n\lambda_1^2) > F_0\} = \alpha_1.$$

Hence $F_0 = F_{\alpha_1}(1+n\lambda_1^2)$, F_{α_1} being the upper $100\alpha_1\%$ point of the F distribution with $\nu_1 = k-1$, $\nu_2 = k(n-1)$.

If condition (*b*) is also to be satisfied we must have

$$\Pr.\{F(1+n\lambda_2^2) > F_0\} > 1-\alpha_2,$$

i.e.

$$\Pr.\{F(1+n\lambda_2^2) > F_{\alpha_1}(1+n\lambda_1^2)\} > 1-\alpha_2.$$

This means that

$$\frac{F_{\alpha_1}}{F_{1-\alpha_2}} < \frac{1+n\lambda_2^2}{1+n\lambda_1^2}, \tag{5}$$

$F_{1-\alpha_2}$ being the lower $100\alpha_2\%$ point of the F distribution with $\nu_1 = k-1$, $\nu_2 = k(n-1)$. The choice of λ_1, λ_2, α_1 and α_2 will depend on practical requirements. Once these are decided upon, the number of groups, k, and the number of samples per group, n, should be chosen so that (5) is satisfied, if possible.

Imagine k fixed, and consider the effect of increasing n. As n increases the ratio $F_{\alpha_1}/F_{1-\alpha_2}$ decreases steadily, approaching the limit $\chi^2_{\alpha_1}/\chi^2_{1-\alpha_2}$ as n approaches infinity. ($\chi^2_{\alpha_1}$ and $\chi^2_{1-\alpha_2}$ represent the upper $100\alpha_1\%$ and lower $100\alpha_2\%$ points respectively of the χ^2 distribution

with $(k-1)$ degrees of freedom.) As n increases the ratio $(1+n\lambda_2^2)/(1+n\lambda_1^2)$ increases steadily from 1 to $(\lambda_2/\lambda_1)^2$. If $\chi^2_{\alpha_1}/\chi^2_{1-\alpha_2}$ is less than $(\lambda_2/\lambda_1)^2$ there will be a number n_0 such that if $n \geqslant n_0$ condition (5) is satisfied. If $\chi^2_{\alpha_1}/\chi^2_{1-\alpha_2}$ is greater than $(\lambda_2/\lambda_1)^2$, on the other hand, it will not be possible to satisfy (5), however large n may be. $\chi^2_{\alpha_1}/\chi^2_{1-\alpha_2}$ is a decreasing function of k, approaching 1 as k tends to infinity. There will therefore be a minimum number of groups, k_0, below which it is impossible to satisfy (5). A short table of such minimum values is given below.

$\alpha_1 = \alpha_2 = 0{\cdot}05$		$\alpha_1 = \alpha_2 = 0{\cdot}01$	
λ_2/λ_1	k_0	λ_2/λ_1	k_0
1·5	35	1·5	68
2	14	2	25
2·5	9	2·5	16
3	7	3	12

It may be noted that if it is reasonable to assume that σ is constant, λ_2/λ_1 is the ratio of the 'unacceptable' to the 'acceptable' limit of between-groups variability.

8. The alternative systems developed in §§ 2 and 3 are of use in the interpretation of the analysis of variance for data arranged according to a cross-classification. We shall consider only the simple case of a two-way cross-classification into l 'rows' and m 'columns', there being n observations in each cell. The symbol x_{ijt} will be used to denote the tth observation in the cell belonging to the ith row and the jth column. The analysis of variance table is:

Source	Sum of squares	Degrees of freedom
Between rows	$mn\sum_{i=1}^{l}(\bar{x}_{i..}-\bar{x}_{...})^2$	$l-1$
Between columns	$ln\sum_{j=1}^{m}(\bar{x}_{.j.}-\bar{x}_{...})^2$	$m-1$
Interaction	$n\sum_{i=1}^{l}\sum_{j=1}^{m}(\bar{x}_{ij.}-\bar{x}_{i..}-\bar{x}_{.j.}+\bar{x}_{...})^2$	$(l-1)(m-1)$
Within cells	$\sum_{i=1}^{l}\sum_{j=1}^{m}\sum_{t=1}^{n}(x_{ijt}-\bar{x}_{ij.})^2$	$lmn-1$

where $\bar{x}_{ij.} = n^{-1}\sum_{t=1}^{n}x_{ijt}$, $\bar{x}_{i..} = (mn)^{-1}\sum_{j=1}^{m}\sum_{t=1}^{n}x_{ijt}$, $\bar{x}_{.j.} = (ln)^{-1}\sum_{i=1}^{l}\sum_{t=1}^{n}x_{ijt}$,

$$\bar{x}_{...} = (lmn)^{-1}\sum_{i=1}^{l}\sum_{j=1}^{m}\sum_{t=1}^{n}x_{ijt}.$$

The question now arises—should the ratio of the between-rows mean square to the within-cells mean square be used to test for differences between rows, or should the ratio to the interaction mean square be used? (A similar question arises, of course, in the analysis between columns.) The answer to this question depends on which of the two set-ups (6) and (7), shown below is the more appropriate.

First, extending (1), we have

$$x_{ijt} = A + R_i + C_j + I_{ij} + z_{ijt}. \tag{6}$$

As before the z_{ijt}'s are independent random variables each with expected value zero and standard deviation σ. The parameter A is introduced to simplify the mathematics and represents the overall average level of the character measured. R_i represents the average departure from this level in the ith row; C_j represents the average departure in the jth column. Without loss of generality it may be assumed that

$$\sum_{i=1}^{l} R_i = \sum_{j=1}^{m} C_j = 0.$$

I_{ij} represents the interaction, or departure from the linear set-up, in the cell belonging to the ith row and the jth column. Without loss of generality it may be assumed that

$$\sum_{i=1}^{l} I_{ij} = 0 = \sum_{j=1}^{m} I_{ij}.$$

The expected values of the various mean squares in the analysis of variance table are, under the conditions summarized by (6):

	Mean square	Expected value
(i)	Between rows	$\sigma^2 + mn \sum_{i=1}^{l} R_i^2/(l-1)$
(ii)	Between columns	$\sigma^2 + ln \sum_{j=1}^{m} C_j^2/(m-1)$
(iii)	Interaction	$\sigma^2 + n \sum_{i=1}^{l} \sum_{j=1}^{m} I_{ij}^2/(l-1)(m-1)$
(iv)	Within cells	σ^2

It is clear, therefore, that the ratio of (i) to (iv) should be used to test the hypothesis $R_1 = R_2 = \ldots = R_l = 0$ when set-up (6) is valid. As in the case of set-up (1), Tang's tables can be used to calculate the power of the test to establish significance when there is a true effect. If the interaction mean square were used instead of the within-cells mean square we should risk reaching an inconclusive result when significance could have been established (i.e. there will be an increase in the second kind of error).

Now consider the alternative set-up

$$x_{ijt} = A + R_i + C_j + z'_{ij} + z_{ijt}, \tag{7}$$

formed by replacing the parameters I_{ij} by the random variables z'_{ij}. The z'_{ij}'s are assumed to be independent of each other and of the z_{ijt}'s, each having expected value zero and standard deviation σ'. Under these conditions the expected values of the mean squares in the analysis of variance table are:

	Mean square	Expected value
(i)	Between rows	$\sigma^2 + n\sigma'^2 + mn \sum_{i=1}^{l} R_i^2/(l-1)$
(ii)	Between columns	$\sigma^2 + n\sigma'^2 + ln \sum_{j=1}^{m} C_j^2/(m-1)$
(iii)	Interaction	$\sigma^2 + n\sigma'^2$
(iv)	Within cells	σ^2

It is apparent that if the ratio (i)/(iv) be used to test the hypothesis $R_1 = R_2 = \ldots = R_l = 0$ when set-up (7) is valid, the numerator will tend to be increased by the σ'^2 effect without any corresponding increase in the denominator. There would thus be a bias towards rejection of the null hypothesis (i.e. there will be an increase in the first kind of error). In this case, therefore, the ratio (i)/(iii) would be a preferable criterion.

In the case of a two-way classification, therefore, unlike the case of k groups, the two forms of theoretical set-up lead to different procedures in the analysis of variance. As before, the two systems, in this case (6) and (7), may be regarded as extreme cases. An intermediate set-up of the form

$$x_{ijt} = A + R_i + C_j + I_{ij} + z'_{ij} + z_{ijt}$$

would possibly be a truer reflexion of the practical position. This being so, it may be as well to consider both the ratio with the within-cells mean square and that with the interaction mean square in the denominator. Since the biases introduced by using the inexact tests are in opposite directions it follows that if the verdicts given by both ratios agree, confidence may be placed in the joint decision. Otherwise, closer consideration must be given to the conditions of collection of the data, to decide whether (6) or (7) is the more appropriate set-up.

Considerations similar to those developed above may be applied to more complicated problems in the analysis of variance. As a general rule, if an interaction is represented by a random variable, it is necessary to take it into account when testing interactions of lower order involving the same factors. Otherwise it is not necessary to do so.

9. Randomization theory provides yet a third system of theoretical set-ups. The essential difference between randomization theory and the systems already described is clearly illustrated by comparison of the theoretical set-ups they imply in the case of simple classification by groups.

It is possible to apply randomization theory to such a situation only if it is reasonable to suppose that any individual in the sample could have occurred in any one of the groups. A typical case arises when the effects of a number of treatments are to be investigated. The experimental material is divided into a number of groups and each group is assigned to a particular treatment.

It is supposed that the experimental arrangements are such that each possible arrangement of the N individuals into k groups of size $n_1, n_2, \ldots, n_k$ respectively is equally likely. The process of randomization in selecting the groups is an attempt to make practice consistent with the theory in this respect. The null hypothesis may then be expressed: 'the observed value of the character measured will be the same for any one individual, whatever be the group in which it is placed.' On the null hypothesis, therefore, there are only N possible observed values $u_1, u_2, \ldots, u_N$. Any observed set of values $x_{11}, \ldots, x_{knk}$ is simply a rearrangement of the u_r's. Since x_{ij} is equally likely to have any of the N values $u_1, u_2, \ldots, u_N$, the theoretical set-up on the null hypothesis may be written

$$x_{ij} = y_{ij}, \tag{8}$$

where y_{ij} is a discontinuous random variable with distribution function

$$\Pr.\{y_{ij} = u_r\} = 1/N \quad (r = 1, 2, \ldots, N).$$

It will be noted that all the y_{ij}'s have the same distribution, but they are not independent. If $y_{ij} = u_r$, then, in general, no other of the y_{ij}'s can take the value u_r.

It has been shown (Fisher, 1935; Welch, 1937, 1938; Pitman, 1937) that in the simple cases of classification by groups and cross-classification considered in this paper the distribution of the mean square ratio on the randomization theory may be replaced by the normal theory F distribution with little fear of serious error. In these cases, therefore, randomization of an experiment ensures that critical limits based on normal theory shall be approximately valid.

If we denote $\bar{u} = N^{-1}\sum_{r=1}^{N} u_r$ by A, (8) may be written

$$x_{ij} = A + y'_{ij},$$

where $y'_{ij} = y_{ij} - A$. The y'_{ij}'s are discontinuous, dependent random variables, each having the same distribution with expected value zero. Logically, alternative hypotheses, corresponding to the existence of differences between the groups, could be introduced according with either of the systems described earlier in this paper. Symbolically these would be expressed

$$x_{ij} = A + B_i + y'_{ij}, \tag{9}$$

$$x_{ij} = A + z'_i + y'_{ij}. \tag{10}$$

(9) corresponds to set-up (1), and (10) to (2), so far as the representation of group differences is concerned.

Set-up (1) may be written in the form

$$x_{ij} = A + B_i + z_{ij}. \tag{11}$$

(B_i has been replaced by $A + B_i$, but this is merely a matter of convenience, and does not affect the nature of the set-up.) We also recall that set-up (2) is

$$x_{ij} = A + z'_i + z_{ij}. \tag{12}$$

Studying set-ups (9) to (12) we notice that while the set-ups (1) and (2) differ in the nature of between-groups variation which they specify, randomization theory implies a modification in the distribution of the individual random residual variation, replacing the independent z_{ij}'s by the dependent y'_{ij}'s.

Although four different set-ups are shown above, it seems that (10) would be used but rarely. The use of constant parameters to represent group effects appears to be more consistent with the ideas of randomization theory.

REFERENCES

DANIELS, H. E. (1939). *J. Roy. Statist. Soc. Suppl.* **6**, 186.
EISENHART, C. (1947). *Biometrics*, **3**, 1.
FISHER, R. A. (1935). *The Design of Experiments*. London: Oliver and Boyd.
HSU, P. L. (1941). *Biometrika*, **32**, 62.
PITMAN, E. J. G. (1937). *Biometrika*, **29**, 322.
TANG, P. C. (1938). *Statist. Res. Mem.* **2**, 126.
WELCH, B. L. (1937). *Biometrika*, **29**, 21.
WELCH, B. L. (1938). *Biometrika*, **30**, 149.

AN EXAMINATION AND FURTHER DEVELOPMENT OF A FORMULA ARISING IN THE PROBLEM OF COMPARING TWO MEAN VALUES

By ALICE A. ASPIN, B.A., *Leeds University*

1. Introduction

In a recent paper B. L. Welch (1947) has developed, by a formal process, an expansion applicable to the problem of comparing two mean values. It is the purpose of the present paper (*a*) to extend this expansion to some further terms, (*b*) to investigate the numerical behaviour of the expansion in some particular cases and (*c*) to consider the comparative merits of a rearranged form of the expansion.

If we have two normal populations with true means α_1 and α_2 respectively and true variances σ_1^2 and σ_2^2; and if we have samples of sizes n_1 and n_2 drawn respectively from these populations, yielding sample means $\bar{x}_1$ and $\bar{x}_2$ and sample variances s_1^2 and s_2^2; then $(\bar{x}_1-\bar{x}_2)$ is distributed normally about $(\alpha_1-\alpha_2)$ with variance $(\sigma_1^2/n_1+\sigma_2^2/n_2)$, and s_1^2 and s_2^2 are distributed as $\chi_1^2\sigma_1^2/(n_1-1)$ and $\chi_2^2\sigma_2^2/(n_2-1)$ respectively. Problems of some importance are either to assess the significance of the observed difference $(\bar{x}_1-\bar{x}_2)$, or to calculate as a function of $(\bar{x}_1-\bar{x}_2)$, s_1^2 and s_2^2, limits within which the population difference $(\alpha_1-\alpha_2)$ may be said to lie with a given probability. In his theoretical discussion of this problem Welch has found it convenient to consider it as a particular case of the following more general problem which, formally, is no more difficult to solve.

Suppose η is any population parameter, estimated by an observed quantity y which is normally distributed with variance $\sigma_y^2 = \sum_{i=1}^{k}\lambda_i\sigma_i^2$. Suppose that, in addition, the data provide estimates s_i^2 of the unknown variances σ_i^2 $(i = 1, 2, \ldots, k)$, based on f_i degrees of freedom, and distributed as

$$p(s_i^2)\,ds_i^2 = \frac{1}{\Gamma(\frac{1}{2}f_i)}\left(\frac{f_i s_i^2}{2\sigma_i^2}\right)^{\frac{1}{2}f_i-1}\exp\left[-\frac{1}{2}\frac{f_i s_i^2}{\sigma_i^2}\right]d\left(\frac{f_i s_i^2}{2\sigma_i^2}\right), \tag{1}$$

where the s_i^2 $(i = 1, 2, \ldots, k)$ are all statistically independent of each other and of y. Let $h(s_1^2, s_2^2, \ldots, s_k^2, P)$ be that function of s^2 and P such that the probability is P that $(y-\eta)$ falls short of $h(s_1^2, s_2^2, \ldots, s_k^2, P)$. Then, if we can find $h(s_1^2, s_2^2, \ldots, s_k^2, P)$ in general, the particular application to the case of the comparison of means will follow by setting $k = 2$, $\eta = (\alpha_1-\alpha_2)$, $y = (\bar{x}_1-\bar{x}_2)$; $\lambda_1 = 1/n_1$ and $\lambda_2 = 1/n_2$.

Writing, for convenience, $h(s^2)$ for $h(s_1^2, s_2^2, \ldots, s_k^2, P)$, Welch has shown that the integral equation which $h(s^2)$ must satisfy can be expressed symbolically in the form

$$\Theta I\left\{\frac{h(w)}{\sqrt{(\Sigma\lambda_i\sigma_i^2)}}\right\} = P. \tag{2}$$

The notation used here is that $I(v)$ stands for the integral, from $-\infty$ to v, of the unit normal probability function and

$$\begin{aligned}\Theta &= \exp\left[-\Sigma\sigma_i^2\partial_i^2 - \tfrac{1}{2}\Sigma f_i\log\left(1-\frac{2\sigma_i^2\partial_i^2}{f_i}\right)\right]\\ &= \exp\left[\Sigma\frac{\sigma_i^4\partial_i^4}{f_i}+\tfrac{4}{3}\Sigma\frac{\sigma_i^6\partial_i^6}{f_i^2}+\text{etc.}\right]\\ &= 1+\Sigma\frac{\sigma_i^4\partial_i^4}{f_i}+\left\{\tfrac{4}{3}\Sigma\frac{\sigma_i^6\partial_i^6}{f_i^2}+\frac{1}{2}\left(\Sigma\frac{\sigma_i^4\partial_i^4}{f_i}\right)^2\right\}+\text{etc.},\end{aligned} \tag{3}$$

where ∂_i^r implies repeated differentiation with respect to w_i and subsequent equation of all w_j to σ_j^2.

When the f_i are all large, the solution of (2) is $h_0(w) = \xi\sqrt{(\Sigma\lambda_i w_i)}$, where ξ is the normal deviate such that $I(\xi) = P$. More generally let us write

$$h(w) = h_0(w) + h_1(w) + h_2(w) + \text{etc.}, \tag{4}$$

where $h_1(w)$ includes terms of order $1/f_i$, $h_2(w)$ terms of order $1/f_i^2$, and so on. Further, suppose that we have calculated terms up to and including the order $1/f_i^r$ and wish to obtain an extra term $h_{r+1}(w)$. To the required order we may then write

$$\begin{aligned} I\left\{\frac{h(w)}{\sqrt{(\Sigma\lambda_i\sigma_i^2)}}\right\} &= I\left\{\frac{h_0(w)+\ldots+h_r(w)+h_{r+1}(w)}{\sqrt{(\Sigma\lambda_i\sigma_i^2)}}\right\} \\ &= I\left\{\frac{h_0(w)+\ldots+h_r(w)}{\sqrt{(\Sigma\lambda_i\sigma_i^2)}}\right\} + \frac{h_{r+1}(w)}{\sqrt{(\Sigma\lambda_i\sigma_i^2)}}\, I'\left\{\frac{h_0(w)}{\sqrt{(\Sigma\lambda_i\sigma_i^2)}}\right\}. \end{aligned} \tag{5}$$

When we operate on this with Θ the second term will, to the order involved, need to be treated only with the unit part of (3). Hence (2) will give

$$\Theta I\left\{\frac{h_0(w)+\ldots+h_r(w)}{\sqrt{(\Sigma\lambda_i\sigma_i^2)}}\right\} + \frac{h_{r+1}(\sigma^2)}{\sqrt{(\Sigma\lambda_i\sigma_i^2)}}\, I'(\xi) = I(\xi). \tag{6}$$

This equation gives $h_{r+1}(\sigma^2)$ explicitly. Since $h_{r+1}(w)$ is the same function of w_i and $h_{r+1}(s^2)$ the same function of s_i^2 as $h_{r+1}(\sigma^2)$ is of σ_i^2, we have therefore an explicit method of deriving successive terms in an expansion for $h(s^2)$.

Pursuing the symbolic method a stage further we may expand $I\left\{\frac{h_0(w)+\ldots+h_r(w)}{\sqrt{(\Sigma\lambda_i\sigma_i^2)}}\right\}$ formally in a Taylor series, thus

$$I\left\{\frac{h_0(w)+\ldots+h_r(w)}{\sqrt{(\Sigma\lambda_i\sigma_i^2)}}\right\} = \exp\left[\left\{\frac{h_0(w)+\ldots+h_r(w)}{\sqrt{(\Sigma\lambda_i\sigma_i^2)}} - \xi\right\} D\right] I(v), \tag{7}$$

where D^r denotes repeated differentiation with respect to v and subsequent equation of v to ξ. The operation Θ in (6) may then be regarded as acting first on the exponential on the right-hand side of (7), producing an expression involving D which must in turn be operated on $I(v)$. Following out this procedure, Welch (1947, p. 31) has obtained the following expressions as far as order $1/f_i^2$:

$$h_0(s^2) = \xi\sqrt{(\Sigma\lambda_i s_i^2)}, \quad h_1(s^2) = \xi\sqrt{(\Sigma\lambda_i s_i^2)}\left\{\frac{(1+\xi^2)}{4}\frac{\left(\Sigma\frac{\lambda_i^2 s_i^4}{f_i}\right)}{(\Sigma\lambda_i s_i^2)^2}\right\},$$

$$h_2(s^2) = \xi\sqrt{(\Sigma\lambda_i s_i^2)}\left[-\frac{(1+\xi^2)}{2}\frac{\left(\Sigma\frac{\lambda_i^2 s_i^4}{f_i^2}\right)}{(\Sigma\lambda_i s_i^2)^2} + \frac{(3+5\xi^2+\xi^4)}{3}\frac{\left(\Sigma\frac{\lambda_i^3 s_i^6}{f_i^2}\right)}{(\Sigma\lambda_i s_i^2)^3} - \frac{(15+32\xi^2+9\xi^4)}{32}\frac{\left(\Sigma\frac{\lambda_i^2 s_i^4}{f_i}\right)^2}{(\Sigma\lambda_i s_i^2)^4}\right]. \tag{8}$$

If we introduce the notation

$$V_{rs} = \frac{\left(\Sigma\frac{\lambda_i^r s_i^{2r}}{f_i^s}\right)}{(\Sigma\lambda_i s_i^2)^r} \tag{9}$$

equations (8) may be more compactly written as

$$\begin{aligned} &h_0(s^2) = \xi\sqrt{(\Sigma\lambda_i s_i^2)}, \quad \frac{h_1(s^2)}{h_0(s^2)} = \tfrac{1}{4}(1+\xi^2)V_{21}, \\ &\frac{h_2(s^2)}{h_0(s^2)} = [-\tfrac{1}{2}(1+\xi^2)V_{22} + \tfrac{1}{3}(3+5\xi^2+\xi^4)V_{32} - \tfrac{1}{32}(15+32\xi^2+9\xi^4)V_{21}^2]. \end{aligned} \tag{10}$$

2. The development of further terms

To find $h_3(s^2)$ we must first extend equation (3) to include terms of order $1/f_i^3$. This gives

$$\boldsymbol{\Theta} = 1+\left(\Sigma\frac{\sigma_i^4\partial_i^2}{f_i}\right)+\left\{\frac{4}{3}\left(\Sigma\frac{\sigma_i^6\partial_i^3}{f_i^2}\right)+\frac{1}{2}\left(\Sigma\frac{\sigma_i^4\partial_i^2}{f_i}\right)^2\right\}+\left\{2\left(\Sigma\frac{\sigma_i^8\partial_i^4}{f_i^3}\right)+\frac{4}{3}\left(\Sigma\frac{\sigma_i^4\partial_i^2}{f_i}\right)\left(\Sigma\frac{\sigma_i^6\partial_i^3}{f_i^2}\right)+\frac{1}{6}\left(\Sigma\frac{\sigma_i^4\partial_i^2}{f_i}\right)^3\right\}. \tag{11}$$

This expression must operate on the exponential

$$\begin{aligned}\exp\left[\left\{\frac{h_0(w)+\ldots+h_2(w)}{\sqrt{(\Sigma\lambda_i\sigma_i^2)}}-\xi\right\}D\right] &= \exp\left[\left\{\frac{h_0(w)}{\sqrt{(\Sigma\lambda_i\sigma_i^2)}}-\xi\right\}D\right]\exp\left[\left\{\frac{h_1(w)+h_2(w)}{\sqrt{(\Sigma\lambda_i\sigma_i^2)}}\right\}D\right]\\ &= \exp\left[\left\{\sqrt{\left(\frac{(\Sigma\lambda_i w_i)}{(\Sigma\lambda_i\sigma_i^2)}\right)}-1\right\}\xi D\right]\left[1+\frac{h_1(w)\,D}{\sqrt{(\Sigma\lambda_i\sigma_i^2)}}+\left\{\frac{h_2(w)\,D}{\sqrt{(\Sigma\lambda_i\sigma_i^2)}}+\frac{1}{2}\frac{h_1^2(w)\,D^2}{(\Sigma\lambda_i\sigma_i^2)}\right\}\right.\\ &\quad\left.+\left\{\frac{h_1(w)\,h_2(w)\,D^2}{(\Sigma\lambda_i\sigma_i^2)}+\frac{1}{6}\frac{h_1^3(w)\,D^3}{(\Sigma\lambda_i\sigma_i^2)^{\frac{3}{2}}}\right\}\right].\end{aligned} \tag{12}$$

$h_1(w)$ and $h_2(w)$ are already known and can be substituted in here. When this is done the successive differentiations required by (11) can be carried out and the first term on the left-hand side of (6) evaluated for the case $(r+1)=3$. This equation then gives $h_3(\sigma^2)$. The algebra involved is heavy and the details will not be given here. The eventual result is found to be

$$\begin{aligned}\frac{h_3(s^2)}{h_0(s^2)} = [&(1+\xi^2)V_{23}-2(3+5\xi^2+\xi^4)V_{33}+\tfrac{1}{8}(15+32\xi^2+9\xi^4)V_{22}V_{21}\\ &+\tfrac{1}{8}(75+173\xi^2+63\xi^4+5\xi^6)V_{43}-\tfrac{1}{12}(105+298\xi^2+140\xi^4+15\xi^6)V_{32}V_{21}\\ &+\tfrac{1}{384}(945+3169\xi^2+1811\xi^4+243\xi^6)V_{21}^3].\end{aligned} \tag{13}$$

The labour involved in calculating groups of terms of successive orders increases rapidly with r. This is due to the increasing number of differential operations introduced with each new term in the expansion of $\boldsymbol{\Theta}$ and also to the rapidly increasing complexity of the expression on which $\boldsymbol{\Theta}$ operates. Hence, even for $(r+1)=3$, a very large number of differentiations with respect to the w_i have to be carried out before the final equation of all w_j to σ_j^2. In extending the work still further to the terms of order $1/f_i^4$, the present author found it necessary to cover over 100 pages with algebra, details of which need not now be given. The eventual result is found to be

$$\begin{aligned}\frac{h_4(s^2)}{h_0(s^2)} = [&-2(1+\xi^2)V_{24}+\tfrac{28}{3}(3+5\xi^2+\xi^4)V_{34}\\ &-\tfrac{1}{4}(15+32\xi^2+9\xi^4)\{V_{23}V_{21}+\tfrac{1}{2}V_{22}^2\}\\ &-\tfrac{3}{2}(75+173\xi^2+63\xi^4+5\xi^6)V_{44}\\ &+\tfrac{1}{2}(105+298\xi^2+140\xi^4+15\xi^6)\{\tfrac{1}{3}V_{22}V_{32}+V_{21}V_{33}\}\\ &+\tfrac{1}{4}(15+33\xi^2+11\xi^4+\xi^6)V_{44}\\ &+\tfrac{1}{5}(735+2170\xi^2+1126\xi^4+168\xi^6+7\xi^8)V_{54}\\ &-\tfrac{1}{64}(945+3169\xi^2+1811\xi^4+243\xi^6)V_{42}V_{21}^2\\ &-\tfrac{1}{18}(945+3354\xi^2+2166\xi^4+425\xi^6+25\xi^8)V_{32}^2\\ &-\tfrac{1}{32}(4725+16586\xi^2+10514\xi^4+1974\xi^6+105\xi^8)V_{21}V_{43}\\ &+\tfrac{1}{96}(10395+42429\xi^2+31938\xi^4+7335\xi^6+495\xi^8)V_{32}V_{21}^2\\ &-\tfrac{1}{6144}(135135+626144\xi^2+542026\xi^4+145320\xi^6+11583\xi^8)V_{21}^4].\end{aligned} \tag{14}$$

3. CHECKS

As so much heavy algebra has been involved in reaching these results, greater confidence will be placed in them if some independent checks are available. One possibility is to try to find an expansion of some function of $h(s^2)$ rather than of $h(s^2)$ itself. It happens, indeed, that the square $h^2(s^2)$ can be developed in terms of successive orders in $1/f_i^r$ by a similar method to that used above without involving quite as much labour. In this development the Θ of equation (2), instead of operating on a normal probability integral, operates on the integral of the distribution of the square of a unit normal deviate. The first approximation to $h^2(s^2)$ is $\xi^2(\Sigma\lambda_i s_i^2)$ so that, analogously to equation (12), we find that Θ has to operate on the product of two factors of which the first is $\exp\left[\left\{\frac{(\Sigma\lambda_i w_i)}{(\Sigma\lambda_i\sigma_i^2)}-1\right\}\xi^2 D\right]$. Since this does not contain a square-root sign, some of the labour of operating with Θ is lightened. A full check of equation (13) using this alternative method has been carried out.

A full independent check of the terms of order $1/f_i^4$ has not been similarly obtained, but a check which is almost as satisfactory is obtained by putting $k = 1$. We then find that (14) reduces to

$$\frac{h_4(s^2)}{h_0(s^2)} = \frac{(-945-1920\xi^2+1482\xi^4+776\xi^6+79\xi^8)}{360\times 4^4 f^4}. \tag{15}$$

This agrees, as it should, with the term in $1/f^4$ in the expansion of the straightforward 'Student' deviate given by R. A. Fisher (1941, p. 151). It is difficult to imagine algebraical mistakes which could have been made in reaching (14) without invalidating this agreement in the particular case, $k = 1$.

4. A REARRANGEMENT OF THE EXPANSION

It will be seen on inspection of equations (10), (13) and (14) that $h_1(s^2)$, $h_2(s^2)$, $h_3(s^2)$ and $h_4(s^2)$ all contain a term having $(1+\xi^2)$ as a factor. Grouping these terms together, the total contribution to $h(s^2)$ from this source is

$$\tfrac{1}{4}\{\xi\sqrt{(\Sigma\lambda_i s_i^2)}\}(1+\xi^2)\{V_{21}-2V_{22}+4V_{23}-8V_{24}\ldots\}. \tag{16}$$

Further, we have

$$\begin{aligned}(\Sigma\lambda_i s_i^2)^2\{V_{21}-2V_{22}+4V_{23}-8V_{24}\ldots\} &= \left(\Sigma\frac{\lambda_i^2 s_i^4}{f_i}\right)-2\left(\Sigma\frac{\lambda_i^2 s_i^4}{f_i^2}\right)+4\left(\Sigma\frac{\lambda_i^2 s_i^4}{f_i^3}\right)-8\left(\Sigma\frac{\lambda_i^2 s_i^4}{f_i^4}\right)\ldots\\ &= \left\{\Sigma\frac{\lambda_i^2 s_i^4}{f_i}\left(1-\frac{2}{f_i}+\frac{4}{f_i^2}-\frac{8}{f_i^3}\ldots\right)\right\}\\ &= \left(\Sigma\frac{\lambda_i^2 s_i^4}{f_i+2}\right).\end{aligned} \tag{17}$$

Hence (16) gives a contribution to $h(s^2)$ which we may denote by $h_1'(s^2)$, where

$$h_1'(s^2) = \xi\sqrt{(\Sigma\lambda_i s_i^2)}\left\{\frac{(1+\xi^2)}{4}\frac{\left(\Sigma\frac{\lambda_i^2 s_i^4}{f_i+2}\right)}{(\Sigma\lambda_i s_i^2)^2}\right\}. \tag{18}$$

Other terms may be combined in a similar manner. Indeed, if we write

$$W_u = \frac{\left(\Sigma\frac{\lambda_i^{u+1} s_i^{2(u+1)}}{(f_i+2)(f_i+4)\ldots(f_i+2u)}\right)}{(\Sigma\lambda_i s_i^2)^{u+1}}, \tag{19}$$

we can rearrange the expansion of $h(s^2)$ in the form

$$h(s^2) = h'_0(s^2) + h'_1(s^2) + h'_2(s^2) + \text{etc.}, \tag{20}$$

where now

$$h'_0(s^2) = \xi\sqrt{(\Sigma\lambda_i s_i^2)}, \quad \frac{h'_1(s^2)}{h'_0(s^2)} = \tfrac{1}{4}(1+\xi^2)\,W_1,$$

$$\frac{h'_2(s^2)}{h'_0(s^2)} = [\tfrac{1}{8}(3+5\xi^2+\xi^4)\,W_2 - \tfrac{1}{32}(15+32\xi^2+9\xi^4)\,W_1^2],$$

$$\frac{h'_3(s^2)}{h'_0(s^2)} = [\tfrac{1}{8}(75+173\xi^2+63\xi^4+5\xi^6)\,W_3 - \tfrac{1}{12}(105+298\xi^2+140\xi^4+15\xi^6)\,W_2W_1 + \tfrac{1}{384}(945+3169\xi^2+1811\xi^4+243\xi^6)\,W_1^3]. \tag{21}$$

The $h'_4(s^2)$ contribution does not come out completely in terms of the W_u's, but involves other expressions in addition. While this contribution cannot on this account be said to be anomalous, the best way of expressing it is not obvious and it is not proposed to enter into a discussion of it here.

Going only as far as terms of order $1/f_i^3$ it will be seen that, by using the W_u's, we have reduced the total number of terms in $h(s^2)$ from eleven to seven. Moreover, the saving in the number of these terms will be more marked as we proceed to higher orders of $1/f_i^4$. In certain circumstances, therefore, particularly if we are computing several probability levels simultaneously, it seems that there might be some gain in using equations (21). However, the question of convergence must be considered before a statement of this kind can properly be made.

5. Numerical investigation

This will be confined to the case $k = 2$ which includes, in particular, the problem of comparing two mean values. It will further be assumed that $f_1 = f_2 = f$, as would happen, for instance, in the mean value problem if the samples drawn from the two populations were equal. (The common f would then be one less than the common sample size.) We are not, however, assuming anything about the relative sizes of the unknown population variances σ_1^2 and σ_2^2.

We shall confine ourselves here to a single probability level $P = 95\,\%$, so that $\xi = 1{\cdot}64485$. Let us then write

$$H_j = 1 + \frac{\left(\sum_{r=1}^{j} h_r(s^2)\right)}{\xi\sqrt{\left(\sum_{i=1}^{k}\lambda_i s_i^2\right)}}, \quad H'_j = 1 + \frac{\left(\sum_{r=1}^{j} h'_r(s^2)\right)}{\xi\sqrt{\left(\sum_{i=1}^{k}\lambda_i s_i^2\right)}}, \tag{22}$$

so that H_j and H'_j are the successive approximations to $h(s^2)/\xi\sqrt{(\Sigma\lambda_i s_i^2)}$, according as the original expansion or the rearranged form is used. The H's depend on the observed data only through the ratio s_1^2/s_2^2 or, equivalently, the ratio $\lambda_1 s_1^2/(\lambda_1 s_1^2 + \lambda_2 s_2^2)$. Numerical values in Table 1 are given against this latter ratio for $f = 6$, 12 and 18. The argument of $\lambda_1 s_1^2/(\lambda_1 s_1^2 + \lambda_2 s_2^2)$ is by tenths from 1·0 to 0·5 (the values 0·4 to 0·0 follow by symmetry and are not therefore shown).

Considering first the expansion in powers of $1/f_i$, the degree of accuracy obtainable is seen from a comparison of H_3 and H_4. When $\lambda_1 s_1^2/(\lambda_1 s_1^2 + \lambda_2 s_2^2)$ equals 1 (or 0) we appear to have four decimal places accurate when $f = 6$, and five decimal places when $f = 12$ or more. In this extreme case $h(s^2)/\xi\sqrt{(\Sigma\lambda_i s_i^2)}$ is a simple 'Student' deviate divided by ξ. We already know, of course, that the series representation of such a quantity converges rapidly (e.g.

Fisher, 1941). In the final column of our Table 1 we give for comparison the values of t/ξ calculated from Mrs M. Merrington's table of the percentage points of 'Student's' t (Merrington, 1942).

In the remainder of Table 1 we have no similar independent source against which we can make any checks and can only be guided by the relative sizes of the H_j. It will be seen that, as we move towards the middle of the range of $\lambda_1 s_1^2/(\lambda_1 s_1^2+\lambda_2 s_2^2)$, the difference between H_3 and H_4 tends to increase. As a result it appears that the accuracy available at $\lambda_1 s_1^2/(\lambda_1 s_1^2+\lambda_2 s_2^2)$ equal to $\frac{1}{2}$ is two decimal places for $f = 6$, three decimals for $f = 12$ and four decimals for $f = 18$. The accuracy improves towards each end of the range of $\lambda_1 s_1^2/(\lambda_1 s_1^2+\lambda_2 s_2^2)$.

Table 1. *Successive approximations to* $h(s^2)/\xi\sqrt{(\Sigma\lambda_i s_i^2)}$

($k = 2$, $f_1 = f_2 = f$, $P = 95\,\%$, $\xi = 1{\cdot}64485$)

$\frac{\lambda_1 s_1^2}{\lambda_1 s_1^2+\lambda_2 s_2^2}$	$f = 6$							
	H_1	H_2	H_3	H_4	H_1'	H_2'	H_3'	t_P/ξ
1·0	1·1544	1·1784	1·1812	1·1814	1·1158	1·1334	1·1266	1·1814
0·9	1·1266	1·1478	1·1524	1·1531	1·0949	1·1125	1·1101	—
0·8	1·1050	1·1176	1·1217	1·1248	1·0787	1·0926	1·0930	—
0·7	1·0895	1·0924	1·0917	1·0934	1·0672	1·0765	1·0771	—
0·6	1·0803	1·0761	1·0698	1·0668	1·0602	1·0659	1·0656	—
0·5	1·0772	1·0703	1·0616	1·0559	1·0579	1·0623	1·0615	—

$\frac{\lambda_1 s_1^2}{\lambda_1 s_1^2+\lambda_2 s_2^2}$	$f = 12$							
	H_1	H_2	H_3	H_4	H_1'	H_2'	H_3'	t_P/ξ
1·0	1·07720	1·08319	1·08354	1·08355	1·06617	1·07496	1·07600	1·0836
0·9	1·06330	1·06861	1·06919	1·06923	1·05426	1·06221	1·06356	—
0·8	1·05249	1·05564	1·05617	1·05636	1·04500	1·05111	1·05242	—
0·7	1·04478	1·04552	1·04544	1·04555	1·03838	1·04253	1·04339	—
0·6	1·04014	1·03908	1·03829	1·03810	1·03441	1·03713	1·03747	—
0·5	1·03859	1·03687	1·03578	1·03542	1·03308	1·03528	1·03541	—

$\frac{\lambda_1 s_1^2}{\lambda_1 s_1^2+\lambda_2 s_2^2}$	$f = 18$							
	H_1	H_2	H_3	H_4	H_1'	H_2'	H_3'	t_P/ξ
1·0	1·05147	1·05413	1·05423	1·05423	1·04632	1·05130	1·05219	1·0542
0·9	1·04220	1·04456	1·04473	1·04474	1·03799	1·04238	1·04324	—
0·8	1·03500	1·03640	1·03655	1·03659	1·03150	1·03485	1·03557	—
0·7	1·02985	1·03018	1·03015	1·03017	1·02687	1·02915	1·02961	—
0·6	1·02676	1·02629	1·02606	1·02602	1·02409	1·02561	1·02584	—
0·5	1·02573	1·02497	1·02465	1·02458	1·02316	1·02440	1·02451	—

Turning now to the rearranged form of the expansion we notice immediately that at the beginning of the table H'_3 does not compare favourably with H_3, since the value of H'_3 is not close to t/ξ.

As we leave this end of the table the numerical differences between H'_3 and H_3 decrease. It appears moreover, from an examination of the relative sizes of H'_1, H'_2 and H'_3, that in the vicinity of the value of $\lambda_1 s_1^2/(\lambda_1 s_1^2+\lambda_2 s_2^2)$ equal to $\frac{1}{2}$, the rearranged expansion converges more quickly than the original one. There does not, however, seem to be a sufficiently strong case for systematically abandoning the original expansion everywhere in favour of the rearranged form, even although the latter involves fewer terms.

The present numerical investigation is a preliminary one and other possible rearrangements of the expansion are being considered before any general suggestions for computing tables of $h(s^2)$ are finally made. It will, perhaps, best summarize the work to the point reached if

Table 2. *Values of* $h(s^2)/\sqrt{(\lambda_1 s_1^2+\lambda_2 s_2^2)}$ *and equivalent* F

($k = 2$; $P = 95\,\%$)

$\frac{\lambda_1 s_1^2}{\lambda_1 s_1^2+\lambda_2 s_2^2}$	$h(s^2)/\sqrt{(\lambda_1 s_1^2+\lambda_2 s_2^2)}$ for			Equivalent F for		
	$f = 6$	$f = 12$	$f = 18$	$f = 6$	$f = 12$	$f = 18$
1·0	1·9432	1·7823	1·7341	6	12	18
0·9	1·897	1·7587	1·7184	6·9	14·3	21·7
0·8	1·85	1·738	1·7050	8·3	17·4	26·3
0·7	1·80	1·720	1·6945	11	21	32
0·6	1·75	1·708	1·6877	15	25	37
0·5	1·74	1·703	1·6853	17	27	39

we present again compactly in Table 2 the final result, H_4, for all the cases considered. A slight modification may be made at this stage, by multiplying through by ξ so that the quantity tabled is now $h(s^2)/\sqrt{(\Sigma\lambda_i s_i^2)}$. The quantities in the first line of the table are then the 'Student' deviates for $f = 6$, 12 and 18. Only as many decimal places are given as appear to be justified by the behaviour of the successive terms in the expansion.

Although $h(s^2)/\sqrt{(\Sigma\lambda_i s_i^2)}$ cannot, beyond the first line of the table, be derived by noting that some quantity follows a simple 'Student' distribution, it is nevertheless of interest to consider what degrees of freedom F, say, a 'Student' deviate would have to possess if it were to have the same percentage points as the values of $h(s^2)/\sqrt{(\Sigma\lambda_i s_i^2)}$ given. Accordingly we have shown these equivalent F's in the last three columns. These values of F enable one, perhaps, to appreciate better the trend of the figures given in the earlier columns.

6. Application to the comparison of two means

The quantity $h(s^2)$ was defined in the first section to satisfy the relation

$$\Pr.[(y-\eta) < h(s^2)] = P, \tag{23}$$

$h(s^2)$ is, of course, a function of P as well as of $s_1^2, s_2^2, \ldots, s_k^2$ and should, perhaps, have been denoted by $h(s^2, P)$, but $h(s^2)$ was used throughout to save a little trouble in writing. The

dependence on P is understood. Turning now to the particular case of comparing mean values we must, as has already been noted, set

$$y = (\bar{x}_1 - \bar{x}_2), \quad \eta = (\alpha_1 - \alpha_2), \quad \lambda_1 = \frac{1}{n_1}, \quad \lambda_2 = \frac{1}{n_2}, \quad f_1 = (n_1 - 1), \quad f_2 = (n_2 - 2), \quad (24)$$

where n_1 and n_2 are the respective sample sizes. Then (23) becomes

$$\text{Pr.}\,[\{(\bar{x}_1 - \bar{x}_2) - (\alpha_1 - \alpha_2)\} < h(s^2)] = P. \quad (25)$$

If we write

$$v = \frac{\{(\bar{x}_1 - \bar{x}_2) - (\alpha_1 - \alpha_2)\}}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}, \quad (26)$$

(25) becomes

$$\text{Pr.}\left[v < \frac{h(s^2)}{\sqrt{(\lambda_1 s_1^2 + \lambda_2 s_2^2)}}\right] = P. \quad (27)$$

The numerator of v is normally distributed about zero with variance $(\sigma_1^2/n_1 + \sigma_2^2/n_2)$. The quantity under the square-root in the denominator of v is an unbiased estimate of $(\sigma_1^2/n_1 + \sigma_2^2/n_2)$ since s_1^2 and s_2^2 are respectively unbiased estimates of σ_1^2 and σ_2^2. The ratio v is, in general, to be distinguished from the quantity

$$u = \frac{\{(\bar{x}_1 - \bar{x}_2) - (\alpha_1 - \alpha_2)\}}{\sqrt{\left\{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}}}, \quad (28)$$

which would be the appropriate one to use if it could be assumed that $\sigma_1 = \sigma_2 = \sigma$. The ratio u would then be referred to the t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom. If u is referred to the t-distribution when, in fact $\sigma_1 \neq \sigma_2$ we are liable to be led into error, as has been shown by Welch (1938). These errors are not serious if the sample sizes are equal $(n_1 = n_2 = n)$ for then u and v are the same quantity. Such error as there is is then due to referring u (or v) to the 'Student' distribution with $2f = 2(n-1)$ degrees of freedom when, in fact, as we have seen, the appropriate percentage point depends to some extent on the observed ratio s_1/s_2. The critical value of v will be read off from a table like Table 2, entering with $f = (n-1)$ and with the ratio of s_1^2/n to $(s_1^2/n + s_2^2/n)$.

To take a particular example, suppose that $n_1 = n_2 = 7$ and quantities $\bar{x}_1$, $\bar{x}_2$, s_1 and s_2 are observed such that

$$\bar{x}_1 - \bar{x}_2 = 3{\cdot}7, \quad \sqrt{(s_1^2/7 + s_2^2/7)} = 1{\cdot}2, \quad \frac{s_1^2/7}{s_1^2/7 + s_2^2/7} = 0{\cdot}64. \quad (29)$$

Then entering Table 2 with $f = 6$ and the ratio 0·64, we have

$$h(s^2)/\sqrt{(\lambda_1 s_1^2 + \lambda_2 s_2^2)} = 1{\cdot}77.$$

If now we wished to test whether the data were consistent with the hypothesis that $\alpha_1 = \alpha_2$, we could compare the numerical value of the ratio $3{\cdot}7/1{\cdot}2 = 3{\cdot}1$ with the significance level 1·77; this, it exceeds greatly. On the other hand, if we wished to compute limits within which $(\alpha_1 - \alpha_2)$ lies with given probability, we would have from (27)

$$\text{Pr.}\,[\{3{\cdot}7 - (\alpha_1 - \alpha_2)\}/1{\cdot}2 < 1{\cdot}77] = 0{\cdot}95. \quad (30)$$

Whence, on rearrangement of the inequality

$$\text{Pr.}\,[(\alpha_1 - \alpha_2) > 3{\cdot}7 - 2{\cdot}12 = 1{\cdot}58] = 0{\cdot}95. \quad (31)$$

Owing to symmetry, the corresponding value of $h(s^2)/\sqrt{(\lambda_1 s_1^2+\lambda_2 s_2^2)}$ for $P = 5\,\%$ will be *minus* 1·77. Thus analogously to (30)

$$\Pr.\left[\frac{\{3{\cdot}7-(\alpha_1-\alpha_2)\}}{1{\cdot}2} < -1{\cdot}77\right] = 0{\cdot}05, \tag{32}$$

whence

$$\Pr.[(\alpha_1-\alpha_2) > 3{\cdot}7+2{\cdot}12 = 5{\cdot}82] = 0{\cdot}05. \tag{33}$$

Combining this with (31)

$$\begin{aligned}\Pr.[1{\cdot}58 < (\alpha_1-\alpha_2) < 5{\cdot}82] &= 0{\cdot}95-0{\cdot}05\\ &= 0{\cdot}90.\end{aligned} \tag{34}$$

In this equation the limits calculated for $(\alpha_1-\alpha_2)$ are not limits given to us before the samples are drawn, but depend on the observed sample statistics. The probability statement has the meaning usually attached to such statements when the method of inverse probability is not being used. At the earlier stage in equation (30) the same holds true. The numbers 3·7, 1·2 and 1·77 entering into this equation are all functions of the sample statistics and the probability statement is strictly about these functions and not about connexions between three pure numbers regarded as fixed and known to us before the samples were drawn. In principle the present procedure is exactly the same as that already familiar in those problems where the 'Student' t-distribution is applicable. The only difference in detail is that, in a 'Student' problem, the number 1·77 in (30) would be replaced by a tabular 'Student' deviate not dependent on the sample statistics observed. In the problem being considered in the present paper, we obtained the figure 1·77 by entering a table one of whose arguments was $\lambda_1 s_1^2/(\lambda_1 s_1^2+\lambda_2 s_2^2)$, so that the value obtained depended, although not very critically, on the observed ratio s_1/s_2. But the principle is really unaltered and the inversion of the inequality (30) proceeds in precisely the same manner as when we are considering a problem soluble directly by the straightforward 'Student' distribution.

In the present example and in the numerical work of the previous section we have confined ourselves to the case $k = 2$ and then to $f_1 = f_2 = f$. In the problem of comparing two means it is when $f_1 \neq f_2$ that the need for calculations of the kind considered here is most apparent for then, as Welch (1938) has pointed out, the u and v of equations (26) and (28) are not the same criterion and the error involved in the usual reference of u to the t-distribution can be more pronounced. We have considered the $f_1 = f_2$ case here first, merely to simplify calculations in an initial numerical investigation. It is proposed in some future work to discuss numerically the case where $f_1 \neq f_2$ and to compare the merits of different ways in which final tables may then be presented. Some consideration will also be given to the case where more than two population variances and their corresponding estimates are involved, i.e. the general case $k > 2$.

REFERENCES

FISHER, R. A. (1941). The asymptotic approach to Behrens' integral, with further tables for the d test of significance. *Ann. Eugen., Lond.*, **11**, 141–72.

MERRINGTON, MAXINE (1942). Table of percentage points of the t-distribution. *Biometrika*, **32**, 311.

WELCH, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, **29**, 350–62.

WELCH, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, **34**, 28–35.

ON THE POWER FUNCTION OF THE LONGEST RUN AS A TEST FOR RANDOMNESS IN A SEQUENCE OF ALTERNATIVES

By G. BATEMAN, *University College, London*

1. Introduction

It has been suggested that the distribution of the longest run in a sequence of alternatives might be used with advantage in quality control work and allied subjects. Mosteller (1941) considered the case of runs above and below the median for a sample of even size and derived a formula for the probability of getting at least one run of a given length or greater when the number of elements of each of the two kinds is the same. Thus, if the alternatives in the sequence are E and $\bar{E}$, he considered explicitly only the case of $2n$ elements, n of which are E and n of which are $\bar{E}$. It is the purpose of this paper to deal with the slightly more general case of unequal numbers of elements of the two kinds and to consider (i) the distribution of the longest run under the hypothesis of randomness, and (ii) the power function when the alternative hypothesis is that of positive dependence in the sequence both for the simple Markoff chain and when the structure of dependence is more complex. In (ii) the conditional power function technique, as given by David (1947) for the distribution of groups, is used, and the two criteria, length of longest run and number of groups, are compared with respect to the same alternative hypotheses.

2. Distribution of the longest run

It will be assumed that there is a sequence of r elements, r_1 of which are E and r_2 of which are $\bar{E}$, where $r_1+r_2=r$. H_0, the hypothesis to be tested, will be that the elements of the sequence are in a random order, and the criterion used to carry out the test will be the length of the longest run of either E's or $\bar{E}$'s. The total number of sequences which can be formed from the r elements is ${}^rC_{r_1}$; this is the fundamental probability set. In order to pick out from this set the sub-set of sequences containing at least one greatest run of a given length, say g, it will be necessary to consider the partitions of r_1 and of r_2 having k as the greatest part, where $k=1,2,\ldots,g$, and to find the number of ways in which they can be combined to form a sequence with at least one part equal to g and no part greater than g. This may be achieved most simply, perhaps, by considering the different ways in which such partitions of r_1 and r_2 form $2t$ or $2t+1$ groups, where $t=1,2,\ldots,r_1-g+1$ for $r_1 \geqslant r_2$. There will be no loss of generality in assuming $r_1 \geqslant r_2$, and it will be understood in what follows that $r_1 \geqslant r_2$.

Let $f_i(t,k)$, where $i=1,2$, denote the number of compositions* of r_i elements into t parts of which the greatest part contains k elements. This can be expressed in terms of binomial coefficients by using the result that the number of compositions of r_i into t parts, none of which exceeds s in magnitude, is the coefficient of x^{r_i} in the expansion of $(x+x^2+\ldots+x^s)^t$, i.e. of $x^t\left(\frac{1-x^s}{1-x}\right)^t$. This coefficient is $\sum_{j=0}^{t}(-)^j\,{}^tC_j\,{}^{r_i-js-1}C_{t-1}$, and expressing it in terms of the above notation we have

$$\sum_{k\leqslant s} f_i(t,k) = \sum_{j=0}^{t}(-)^j\,{}^tC_j\,{}^{r_i-js-1}C_{t-1}. \qquad (1)$$

* Compositions of a number are merely partitions of a number in which the order is taken into account.

Table 1. *Probability of obtaining at least one greatest run, of either kind of element, of given length, g*

(The probabilities are obtained by dividing the number tabled by the corresponding value of ${}^{r}C_{r_1}$)

				Values of g																	
r	r_1	r_2	${}^{r}C_{r_1}$	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
10	9	1	10	.	.	.	.	.	.	.	.	.	.	2	2	2	2	2	.	.	.	.
	8	2	45	.	.	.	.	.	.	.	.	.	.	.	3	6	9	12	12	3	.	.
	7	3	120	.	.	.	.	.	.	.	.	.	.	.	.	4	12	24	40	36	4	.
	6	4	210	.	.	.	.	.	.	.	.	.	.	.	.	.	5	20	51	91	43	.
	5	5	252	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	48	110	82	2
11	10	1	11	.	.	.	.	.	.	.	.	.	2	2	2	2	2	1	.	.	.	.
	9	2	55	.	.	.	.	.	.	.	.	.	.	3	6	9	12	15	9	1	.	.
	8	3	165	.	.	.	.	.	.	.	.	.	.	.	4	12	24	40	54	30	1	.
	7	4	330	.	.	.	.	.	.	.	.	.	.	.	.	5	20	50	100	107	48	.
	6	5	462	.	.	.	.	.	.	.	.	.	.	.	.	.	6	33	106	204	112	1
12	11	1	12	.	.	.	.	.	.	.	.	2	2	2	2	2	2	.	.	.	.	.
	10	2	66	.	.	.	.	.	.	.	.	.	3	6	9	12	15	15	6	.	.	.
	9	3	220	.	.	.	.	.	.	.	.	.	.	4	12	24	40	60	60	20	.	.
	8	4	495	.	.	.	.	.	.	.	.	.	.	.	5	20	50	100	165	140	15	.
	7	5	792	.	.	.	.	.	.	.	.	.	.	.	.	6	30	92	220	330	114	.
	6	6	924	.	.	.	.	.	.	.	.	.	.	.	.	.	12	72	224	408	206	2
13	12	1	13	.	.	.	.	.	.	.	2	2	2	2	2	2	1	.	.	.	.	.
	11	2	78	.	.	.	.	.	.	.	.	3	6	9	12	15	18	12	3	.	.	.
	10	3	286	.	.	.	.	.	.	.	.	.	4	12	24	40	60	78	58	10	.	.
	9	4	715	.	.	.	.	.	.	.	.	.	.	5	20	50	100	175	230	130	5	.
	8	5	1287	.	.	.	.	.	.	.	.	.	.	.	6	30	90	211	410	453	87	.
	7	6	1716	.	.	.	.	.	.	.	.	.	.	.	.	7	46	177	466	735	284	1

Table 1. (*continued*)

(The probabilities are obtained by dividing the number tabled by the corresponding value of ${}^rC_{r_1}$)

				Values of g																	
r	r_1	r_2	${}^rC_{r_1}$	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
14	13	1	14	.	.	.	.	.	.	2	2	2	2	2	2	2	.	.	.	.	.	.
	12	2	91	.	.	.	.	.	.	.	3	6	9	12	15	18	18	9	1	.	.	.
	11	3	364	.	.	.	.	.	.	.	.	4	12	24	40	60	84	88	48	4	.	.
	10	4	1001	.	.	.	.	.	.	.	.	.	5	20	50	100	175	270	280	100	1	.
	9	5	2002	.	.	.	.	.	.	.	.	.	.	6	30	90	210	420	668	528	50	.
	8	6	3003	.	.	.	.	.	.	.	.	.	.	.	7	42	150	418	919	1167	300	.
	7	7	3432	.	.	.	.	.	.	.	.	.	.	.	.	14	100	378	968	1454	516	2
15	14	1	15	.	.	.	.	.	2	2	2	2	2	2	2	1	.	.	.	.	.	.
	13	2	105	.	.	.	.	.	.	3	6	9	12	15	18	21	15	6	.	.	.	.
	12	3	455	.	.	.	.	.	.	.	4	12	24	40	60	84	106	90	34	1	.	.
	11	4	1365	.	.	.	.	.	.	.	.	5	20	50	100	175	280	370	300	65	.	.
	10	5	3003	.	.	.	.	.	.	.	.	.	6	30	90	210	420	741	960	525	21	.
	9	6	5005	.	.	.	.	.	.	.	.	.	.	7	42	147	394	900	1655	1614	246	.
	8	7	6435	.	.	.	.	.	.	.	.	.	.	.	8	61	270	846	1954	2576	719	1
20	19	1	20	2	2	2	2	2	2	2	2	2	2	.	.	.	.	.	.	.	.	.
	18	2	190	.	3	6	9	12	15	18	21	24	27	27	18	9	1	.	.	.	.	.
	17	3	1140	.	.	4	12	24	40	60	84	112	144	180	196	168	96	20	.	.	.	.
	16	4	4845	.	.	.	5	20	50	100	175	280	420	600	815	960	880	470	70	.	.	.
	15	5	15504	.	.	.	.	6	30	90	210	420	756	1260	1980	2880	3540	3086	1190	56	.	.
	14	6	38760	.	.	.	.	.	7	42	147	392	882	1764	3234	5523	8442	10192	7007	1127	1	.
	13	7	77520	.	.	.	.	.	.	8	56	224	672	1680	3696	7392	13516	20936	21672	7556	112	.
	12	8	125970	.	.	.	.	.	.	.	9	72	324	1080	2973	7180	15708	30124	42164	24945	1391	.
	11	9	167930	.	.	.	.	.	.	.	.	10	90	456	1736	5464	14784	34030	58406	47640	5344	.
	10	10	184756	.	.	.	.	.	.	.	.	.	20	208	1140	4464	13180	34570	64058	58290	8194	2

It follows immediately that

$$f_i(t,s) = \sum_{j=1}^{i} (-)^{j+1}\, {}^{t}C_j \left[{}^{r_1-j(s-1)-1}C_{i-1} - {}^{r_1-js-1}C_{i-1} \right].$$

If $N(2t, g \mid r_1 r_2)$ denotes the number of sequences of $2t$ groups when at least one group contains g elements and no group contains more than g elements, then

$$N(2t, g \mid r_1 r_2) = 2\left[f_1(t,g) \sum_{k \leqslant g} f_2(t,k) + f_2(t,g) \sum_{k \leqslant g-1} f_1(t,k) \right]. \tag{2}$$

The factor 2 is introduced to allow for the fact that the sequence may begin with E or with $\bar{E}$. In the same way it may be seen that

$$\begin{aligned} N(2t+1, g \mid r_1 r_2) = &\left[f_1(t+1,g) \sum_{k \leqslant g} f_2(t,k) + f_2(t,g) \sum_{k \leqslant g-1} f_1(t+1,k) \right] \\ &+ \left[f_1(t,g) \sum_{k \leqslant g} f_2(t+1,k) + f_2(t+1,g) \sum_{k \leqslant g-1} f_1(t,k) \right]. \end{aligned} \tag{3}$$

As it will be required later, $\phi(t_1, t_2, g)$ will be used for the expression

$$f_1(t_1,g) \sum_{k \leqslant g} f_2(t_2,k) + f_2(t_2,g) \sum_{k \leqslant g-1} f_1(t_1,k) \quad \text{provided that} \quad |t_1 - t_2| \leqslant 1.$$

Thus $N(2t, g \mid r_1 r_2) = 2\phi(t,t,g)$ and $N(2t+1, g \mid r_1 r_2) = \phi(t+1,t,g) + \phi(t,t+1,g)$.
The enumeration of the required subset is completed by summing $N(2t, g \mid r_1 r_2)$ and $N(2t+1, g \mid r_1 r_2)$ over all groups, i.e. from $t = 1$ to $t = r_1 - g + 1$. If the number in this subset be denoted by $N(g \mid r_1 r_2)$, then

$$N(g \mid r_1 r_2) = \sum_{t=1}^{r_1-g+1} (2\phi(t,t,g) + \phi(t+1,t,g) + \phi(t,t+1,g)). \tag{4}$$

Hence in a sequence of r elements, r_1 of which are E and r_2 of which are $\bar{E}$ where $r_1 + r_2 = r$ and $r_1 \geqslant r_2$, the probability that the longest run consists of g elements is

$$P\{g \mid r_1 r_2\} = \frac{N(g \mid r_1 r_2)}{{}^{r}C_{r_1}},$$

and the complete probability distribution of the length of the longest run, given r_1 and r_2, is obtained by letting g take all possible values. Values of the function (4) for $r = 10$ to 15, and $r = 20$ are given in Table 1; values of the functions (2) and (3) for $r_1 = 14$, $r_2 = 6$, illustrating the form of the correlation between T, the number of groups, and g, are given in Table 2.

If, as is frequently the case in statistical applications, we require only the probability of the longest run having a length greater than or equal to a given value, say g_0, it will be considerably simpler to calculate $\sum_{g \geqslant g_0} \phi(t_1, t_2, g)$ directly; for using result (1) it follows that

$$\begin{aligned} \sum_{g \geqslant g_0} \phi(t_1,t_2,g) &= {}^{r_1-1}C_{t_1-1}\, {}^{r_2-1}C_{t_2-1} - \prod_{i=1}^{2} \left(\sum_{j=0}^{t_i} (-)^j\, {}^{t_i}C_j\, {}^{r_i-j(g_0-1)-1}C_{t_i-1} \right) && \text{for } |t_1 - t_2| \leqslant 1 \\ &= 0 && \text{for } |t_1 - t_2| > 1, \end{aligned}$$

and for brevity we shall write $\phi(t_1, t_2, g \geqslant g_0)$ for this expression.

The probability that the length of the greatest run, g, is greater than or equal to g_0 is immediate, for

$$\begin{aligned} P\{g \geqslant g_0 \mid r_1 r_2\} &= \sum_{g=g_0}^{r_1} P\{g \mid r_1 r_2\} \\ &= \frac{1}{{}^{r}C_{r_1}} \left[\sum_{t=1}^{r_1-g_0+1} (2\phi(t,t,g \geqslant g_0) + \phi(t+1,t,g \geqslant g_0) + \phi(t,t+1,g \geqslant g_0)) \right]. \end{aligned}$$

In the same way

$$P\{g \leqslant g_0 \,|\, r_1 r_2\} = \frac{1}{{}_rC_{r_1}}\left[\sum_{t=a}^{r_2}\left(2\phi(t,t,g\leqslant g_0)+\phi(t+1,t,g\leqslant g_0)+\phi(t,t+1,g\leqslant g_0)\right)\right],$$

where

$$\phi(t_1,t_2,g\leqslant g_0) = \sum_{g\leqslant g_0}\phi(t_1,t_2,g) = \prod_{i=1}^{2}\left(\sum_{j=0}^{t_i}(-)^j\,{}_{t_i}C_j\,{}_{r_i-jg_0-1}C_{t_i-1}\right) \quad \text{for} \quad |t_1-t_2|\leqslant 1$$

$$= 0 \quad \text{for} \quad |t_1-t_2|>1$$

and

$$a = \left[\frac{r_1+g_0-1}{g_0}\right].$$

Table 2. *The joint distribution of T and g for the case $r_1 = 14$, $r_2 = 6$*

(T = number of groups in the sequence)

T \ g	14	13	12	11	10	9	8	7	6	5	4	3	2	Total
2	2	.	.	.	.	.	.	.	.	.	.	.	.	2
3	5	2	2	2	2	2	2	1	.	.	.	.	.	18
4	.	20	20	20	20	20	20	10	.	.	.	.	.	130
5	.	20	35	50	65	80	95	100	60	15	.	.	.	520
6	.	.	60	120	180	240	300	360	240	60	.	.	.	1560
7	.	.	30	100	210	360	550	780	900	610	100	.	.	3640
8	.	.	.	80	240	480	800	1200	1560	1160	200	.	.	5720
9	.	.	.	20	110	320	700	1300	2140	2590	1350	50	.	8580
10	.	.	.	.	50	200	500	1000	1750	2300	1300	50	.	7150
11	.	.	.	.	5	50	200	550	1225	2255	2410	455	.	7150
12	.	.	.	.	.	12	60	180	420	810	912	180	.	2574
13	.	.	.	.	.	.	7	42	147	392	735	392	1	1716
Total	7	42	147	392	882	1764	3234	5523	8442	10192	7007	1127	1	38760

It may be noted that if longest runs of the E's only are considered then (2) and (3) reduce to $2f_1(t,g)\,{}_{r_2-1}C_{t-1}$ and $f_1(t+1,g)\,{}_{r_2-1}C_{t-1}+f_1(t,g)\,{}_{r_2-1}C_t$ respectively.

If g_E be used to denote the length of the longest run of E's in a sequence, then

$$N\{g_E = s \,|\, r_1 r_2\} = \sum_{t=1}^{r_1-s+1} f_1(t,s)\left[{}_{r_2-1}C_{t-2}+2\,{}_{r_2-1}C_{t-1}+{}_{r_2-1}C_t\right]$$

$$= \sum_{t=1}^{r_1-s+1} f_1(t,s)\,{}_{r_2+1}C_t,$$

and

$$N\{g_E \geqslant s \,|\, r_1 r_2\} = \sum_{t=1}^{r_1-s+1} {}_{r_2+1}C_t \sum_{j=1}^{t}(-)^{j+1}\,{}_tC_j\,{}_{r_1-j(s-1)-1}C_{t-1}.$$

Writing ${}_tC_j\,{}_{r_2+1}C_t = {}_{r_2+1}C_j\,{}_{r_2-j+1}C_{t-j}$, interchanging the order of summation and using the relation that $\sum_{i=0}^{m}{}_mC_{k+i}\,{}_nC_i = {}_{m+n}C_{k+n}$, we have that

$$P\{g_E \geqslant s \,|\, r_1 r_2\} = \frac{1}{{}_rC_{r_1}}\left[\sum_{j=1}^{[r_1/s]}(-)^{j+1}\,{}_{r_2+1}C_j\,{}_{r-js}C_{r_2}\right].$$

This result will be referred to in § 9.

3. Conditional power function for the simple Markoff chain

It is clear that the length of the longest run may be used as a criterion for testing randomness in a sequence in the same way as the number of groups or runs. If a longest run of length g_0 be observed, then, as shown in § 2, the probability of obtaining a run of length greater than

or equal to g_0, or of length less than or equal to g_0, can be found. If either of these probabilities is less than some arbitrarily assigned significance level, then we may reject H_0, the hypothesis of randomness. But the possible alternative hypotheses will be different if we judge on the upper tail (i.e. for g large) or the lower tail (i.e. for g small). When the observed g_0 is so large that the hypothesis H_0 is rejected, then a possible alternative hypothesis, H_1, might be that there is positive dependence of some kind in the sequence; that is to say the elements are not independent and given that E has occurred it becomes more likely that it will occur again. On the other hand if the observed g_0 is small, then a possible alternative hypothesis might be that there is negative dependence in the sequence.

Consider a sequence of possible events $e_1, e_2, \ldots, e_r$ which may or may not be independent; it is well known that

$$P\{e_1 e_2 \ldots e_r\} = \prod_{i=1}^{r} P\{e_i \mid e_1 e_2 \ldots e_{i-1}\}.$$

If the events are independent then the relation reduces to

$$P\{e_1 e_2 \ldots e_r\} = \prod_{i=1}^{r} P\{e_i\}.$$

This is the basis of H_0, the hypothesis of randomness.

If there is dependence as in a simple Markoff chain, each event will be dependent only on the event immediately preceding it, and we shall have

$$P\{e_1 e_2 \ldots e_r\} = \prod_{i=1}^{r} P\{e_i \mid e_{i-1}\}.$$

This is the basis of H_1, the possible alternative hypothesis to H_0, and we shall take into account only the case where the dependence is positive.

Now consider a sequence of r events composed of alternatives E or $\bar{E}$, and let E_i denote the happening of the event E at the ith trial. H_0 and H_1 will be as above with capital letters replacing small ones. Under the binomial hypothesis, i.e. under H_0, we write

$$P\{E_i\} = p \quad \text{and} \quad P\{\bar{E}_i\} = q \quad \text{for} \quad i = 1, 2, \ldots, r, \quad \text{where } p+q = 1.$$

and we shall have

$$P\{g \mid r_1 r_2 H_0\} = \frac{N(g \mid r_1 r_2) p^{r_1} q^{r_2}}{\sum_{\text{all } g} N(g \mid r_1 r_2) p^{r_1} q^{r_2}} = \frac{N(g \mid r_1 r_2)}{{}^{r}C_{r_1}}.$$

The hypothesis, H_0, is rejected if $g \geqslant g_\alpha$ where

$$P\{g \geqslant g_\alpha \mid r_1 r_2 H_0\} = \alpha, \tag{5}$$

and g_α is chosen so that α is as near as possible to the chosen significance level. Such values of g_α are shown in Table 3 for $r = 10, 15, 20$, and α in the neighbourhood of 0·05. This table could clearly be extended if desired.

As a likely alternative to H_0, we take H_1, and suppose that

$$P\{E_1\} = P, \qquad P\{\bar{E}_1\} = Q \quad \text{where} \quad P+Q = 1,$$
$$P\{E_i \mid E_{i-1}\} = P_1, \quad P\{\bar{E}_i \mid E_{i-1}\} = Q_1,$$
$$P\{E_i \mid \bar{E}_{i-1}\} = P_2, \quad P\{\bar{E}_i \mid \bar{E}_{i-1}\} = Q_2 \quad \text{where} \quad P_j + Q_j = 1 \text{ for } j = 1, 2.$$

The conditional power function $P\{g \mid r_1 r_2 H_1\}$ follows in a straightforward way by considering the partitions of r_1 and r_2 as was done to obtain the distribution of g under H_0. For each partition within a given number of $2t$ or $2t+1$ groups the multiplying probabilities are the same, for all that matters is the number of transitions from E to $\bar{E}$ and back again. Thus for a given

sequence of $2t$ groups beginning with E there are $2t-1$ transitions, t from E to $\bar{E}$ and $t-1$ from $\bar{E}$ to E and the remaining r_1-t and r_2-t are permanences of E and $\bar{E}$ respectively. The probability of obtaining a given sequence of $2t$ groups is equal to

$$P P_1^{r_1-t} P_2^{t-1} Q_1^{t} Q_2^{r_2-t} + Q Q_1^{t-1} Q_2^{r_2-t} P_1^{r_1-t} P_2^{t},$$

which may be written

$$\left(\frac{P}{P_2}+\frac{Q}{Q_1}\right)\left(\frac{P_2Q_1}{Q_2P_1}\right)^t P_1^{r_1} Q_2^{r_2}.$$

In the same way the probability of obtaining a given sequence of $t+1$ groups of E and t of $\bar{E}$ is $\dfrac{P}{P_1}\left(\dfrac{P_2Q_1}{Q_2P_1}\right)^t P_1^{r_1} Q_2^{r_2}$, and of t groups of E and $t+1$ of $\bar{E}$ is $\dfrac{Q}{Q_2}\left(\dfrac{P_2Q_1}{Q_2P_1}\right)^t P_1^{r_1} Q_2^{r_2}$.

Table 3. *Values of g_α such that $P\{g \geqslant g_\alpha \mid r_1 r_2\} = \alpha$, where α is as near as possible to* 0·05

(The value of α is given in brackets)

r_1 or r_2 \ r	10	15	20
2	8 (0·067)	13 (0·029)	17 (0·047)
3	7 (0·033)	11 (0·035)	15 (0·035)
4	6 (0·024)	9 (0·055)	12 (0·036)
5	5 (0·040)	8 (0·042)	11 (0·049)
6	—	7 (0·039)	10 (0·038)
7	—	6 (0·053)	9 (0·034)
8	—	—	8 (0·035)
9	—	—	7 (0·046)
10	—	—	7 (0·032)

The joint probability distribution of $2t$ and g is given by

$$P\{2t, g \mid r_1 r_2 H_1\} = \frac{\phi(t,t,g)\left(\dfrac{P}{P_2}+\dfrac{Q}{Q_1}\right)\left(\dfrac{P_2Q_1}{Q_2P_1}\right)^t}{\sum\limits_t \sum\limits_g \left(\dfrac{P_2Q_1}{Q_2P_1}\right)^t \left\{\phi(t,t,g)\left(\dfrac{P}{P_2}+\dfrac{Q}{Q_1}\right)+\phi(t+1,t,g)\dfrac{P}{P_1}+\phi(t,t+1,g)\dfrac{Q}{Q_2}\right\}}$$

$$= \frac{\phi(t,t,g)\left(\dfrac{P}{P_2}+\dfrac{Q}{Q_1}\right)\left(\dfrac{P_2Q_1}{Q_2P_1}\right)^t}{\sum\limits_{t=1}^{r_1} \left(\dfrac{P_2Q_1}{Q_2P_1}\right)^t \left\{{}_{r_1-1}C_{t-1}\,{}_{r_2-1}C_{t-1}\left(\dfrac{P}{P_2}+\dfrac{Q}{Q_1}\right)+{}_{r_1-1}C_t\,{}_{r_2-1}C_{t-1}\dfrac{P}{P_1}+{}_{r_1-1}C_{t-1}\,{}_{r_2-1}C_t\dfrac{Q}{Q_2}\right\}},$$

and similarly

$$P\{2t+1, g \mid r_1 r_2 H_1\} = \frac{\left(\phi(t+1,t,g)\dfrac{P}{P_1}+\phi(t,t+1,g)\dfrac{Q}{Q_2}\right)\left(\dfrac{P_2Q_1}{Q_2P_1}\right)^t}{\sum\limits_{t=1}^{r_1} \left(\dfrac{P_2Q_1}{Q_2P_1}\right)^t \left\{{}_{r_1-1}C_{t-1}\,{}_{r_2-1}C_{t-1}\left(\dfrac{P}{P_2}+\dfrac{Q}{Q_1}\right)+{}_{r_1-1}C_t\,{}_{r_2-1}C_{t-1}\dfrac{P}{P_1}+{}_{r_1-1}C_{t-1}\,{}_{r_2-1}C_t\dfrac{Q}{Q_2}\right\}}.$$

The probability distribution of g is obtained by summing over all t from $t = 1$ to $r_1 - g + 1$, that is to say

$$P\{g \mid r_1 r_2 H_1\} = \frac{\sum_{t=1}^{r_1-g+1} \left\{\phi(t,t,g)\left(\frac{P}{P_2}+\frac{Q}{Q_1}\right)+\phi(t+1,t,g)\frac{P}{P_1}+\phi(t,t+1,g)\frac{Q}{Q_2}\right\}\left(\frac{P_2Q_1}{Q_2P_1}\right)^t}{\sum_{t=1}^{r_2}\left(\frac{P_2Q_1}{Q_2P_1}\right)^t\left\{{}^{r_1-1}C_{t-1}\,{}^{r_2-1}C_{t-1}\left(\frac{P}{P_2}+\frac{Q}{Q_1}\right)+{}^{r_1-1}C_t\,{}^{r_2-1}C_{t-1}\frac{P}{P_1}+{}^{r_1-1}C_{t-1}\,{}^{r_2-1}C_t\frac{Q}{Q_2}\right\}}. \quad (6)$$

For the power of the test we require $P\{g \geqslant g_\alpha \mid r_1 r_2 H_1\}$, where g_α is given by (5). This probability is obtained by summing the left-hand side of (6) from $g = g_\alpha$ to r_1, and in practice it is simpler to sum first with respect to g and then with respect to t using the expression for $\sum_{g=g_\alpha}^{r_1} \phi(t_1, t_2, g)$ given in § 2.

4. Comparison of certain power curves for the criteria, length of longest run and number of groups

Throughout the computation of the power curves it has been assumed that

$$P = PP_1 + QP_2. \quad (7)$$

This condition is arrived at by using the relation $P\{E_i\} = P\{E_{i-1}E_i\} + P\{\bar{E}_{i-1}E_i\}$ on the assumption that $P\{E_i\} = P$ and $P\{\bar{E}_i\} = Q$ for all i; that is to say we are assuming that the probability of the event E occurring at the ith trial when nothing is known about the results of the preceding trials is independent of i. This in effect implies that the start of the sequence of observations is a randomly selected point in a longer sequence following the same law.

The power function for either criterion is given as a function of the three parameters P, P_1 and P_2, but, on application of (7), reduces to a function of two parameters. If we take these as P_1 and P_2, we note that P is a constant, k, on straight lines whose equations are $kP_1 + (1-k)P_2 = k$. The power functions for the number of groups criterion, as plotted by David for $P = 0{\cdot}5$, $0{\cdot}6$ and $0{\cdot}75$, are then sections of the power surface cutting the (P_1, P_2) plane in the straight lines (1), (2) and (3) shown in Fig. 1. As the alternative hypothesis under consideration is $P_1 > P_2$, the curves have been taken as starting at the central diagonal $P_1 = P_2 = P$, that is to say we are not interested in the power for $P_1 < P_2$. Fig. 1 shows contours of the power surface for the number of groups criterion when $r_1 = r_2 = 10$ and the significance level is $0{\cdot}05$. The probability of establishing a difference, using this significance level, when $P_1 > P_2$ is about $0{\cdot}22$, $0{\cdot}56$ and $0{\cdot}89$ respectively on the three contours shown.

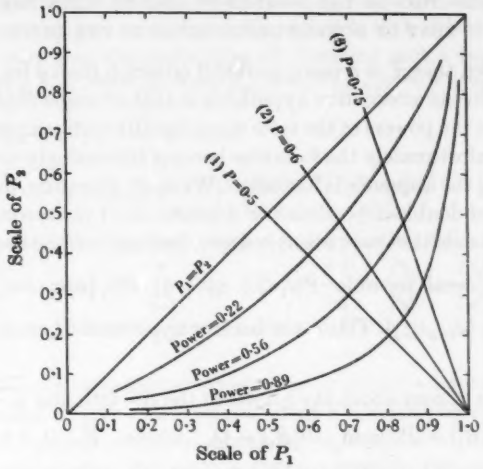
Alternatively we could express the power function in terms of P and $P_1 - P_2$. If, following Markoff (1913, p. 45), we take $P_1 - P_2 = \delta$, then $|\delta|$ is a measure of the degree of dependence in the sequence, and the sign of δ indicates the direction of dependence (+ for positive dependence, − for negative dependence). When $\delta = 0$, the observations are independent. Using (7) it follows that

$$P_1 = P + \delta Q, \quad Q_1 = Q(1-\delta), \quad P_2 = P(1-\delta), \quad Q_2 = Q + \delta P,$$

and the power function for either criterion may be expressed in terms of P and δ. The power curves for the two criteria when the alternative hypothesis is $\delta > 0$ are shown in Fig. 2 for $r_1 = r_2 = 10$, $r_1 = 14$, $r_2 = 6$ and $P = 0{\cdot}5$, $0{\cdot}6$, $0{\cdot}75$.* The chance, α, of rejecting H_0 when it is true is different for the two criteria; this is inevitable owing to the discontinuity of the dis-

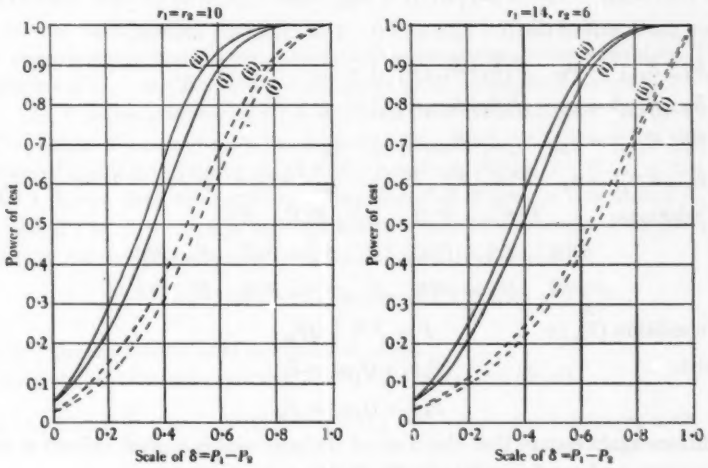
* When $P = 0{\cdot}5$ and $P = 0{\cdot}6$ the power curves are so close as to be indistinguishable on the graphs. This can be seen, too, from the contours in Fig. 1, for the contours are very nearly parallel to the diagonal $\delta = P_1 - P_2 = 0$, for this range of P.

tributions. For example, when $r_1 = 14$, $r_2 = 6$, $P\{g \geqslant 10 \mid H_0\} = 0{\cdot}038$ (see Table 3), while $P\{T \leqslant 6 \mid H_0\} = 0{\cdot}058$. These are the values of α nearest to the 0·05 level in each case. But, allowing for this difference, it is seen on comparing similar curves for the two criteria that T, the number of groups criterion, is the more powerful in detecting departures from randomness of the single dependence kind. This might perhaps have been expected on intuitive grounds.



The 5% significance level has been used.

Fig. 1. Contours of the power surface for the number of groups criterion, T. ($r_1 = r_2 = 10$)



Comparison of power curves for T and g.

Fig. 2. Case where the alternative hypothesis is positive dependence of the simple kind:
— $P\{T \leqslant T_\alpha \mid r_1 r_2 H_1\}$ --- $P\{g \geqslant g_w \mid r_1 r_2 H_1\}$
For curves (i) $P = 0{\cdot}5$ or $0{\cdot}6$ (see footnote to § 4). For curves (ii) $P = 0{\cdot}75$.

It can further be noted that for either criterion the power is greater when $r_1 = r_2$, though the difference is more marked for the length of the longest run. In quality control work it is usual to consider runs above and below the median rather than above and below the mean. This ensures that $r_1 = r_2$ and gives increased power to the test.

5. The distribution of the number of groups when the hypothesis is that of double dependence in the chain

In § 4 it has been shown that T is a more powerful criterion than g for detecting departures from randomness when the alternative hypothesis is that of single dependence. It would be interesting to compare the powers of the tests when the alternative hypothesis is dependence of a different kind. Unfortunately the formulae become increasingly complex as the number of parameters defining the sequence is increased. We shall, therefore, deal only with the case of what could be called double dependence or dependence of the second order.

It is assumed in this case that each event is dependent only on the two events immediately preceding it. The general formula $P\{e_1 e_2 \dots e_r\} = \prod_{i=1}^{r} P\{e_i \mid e_1 e_2 \dots e_{i-1}\}$ then reduces to $P\{e_1 e_2 \dots e_r\} = \prod_{i=1}^{r} P\{e_i \mid e_{i-2} e_{i-1}\}$. This is the basis of hypothesis H_2, and it is further assumed under H_2 that

(i) when nothing is known about the results of the $(i-2)$th and $(i-1)$th trials,

$$P\{E_i\} = P \quad \text{and} \quad P\{\bar{E}_i\} = Q; \quad \text{where} \quad P+Q = 1;$$

(ii) when the result of the $(i-1)$th trial is known, but not the result of the $(i-2)$th trial,

$$P\{E_i \mid E_{i-1}\} = P_1, \quad P\{\bar{E}_i \mid E_{i-1}\} = Q_1,$$

$$P\{E_i \mid \bar{E}_{i-1}\} = P_2, \quad P\{\bar{E}_i \mid \bar{E}_{i-1}\} = Q_2, \quad \text{where} \quad P_j + Q_j = 1 \text{ for } j = 1, 2;$$

(iii) when the result of the $(i-2)$th and $(i-1)$th trials are known,

$$P\{E_i \mid E_{i-2} E_{i-1}\} = p_1, \quad P\{\bar{E}_i \mid E_{i-2} E_{i-1}\} = q_1,$$

$$P\{E_i \mid E_{i-2} \bar{E}_{i-1}\} = p_2, \quad P\{\bar{E}_i \mid E_{i-2} \bar{E}_{i-1}\} = q_2,$$

$$P\{E_i \mid \bar{E}_{i-2} E_{i-1}\} = p_3, \quad P\{\bar{E}_i \mid \bar{E}_{i-2} E_{i-1}\} = q_3,$$

$$P\{E_i \mid \bar{E}_{i-2} \bar{E}_{i-1}\} = p_4, \quad P\{\bar{E}_i \mid \bar{E}_{i-2} \bar{E}_{i-1}\} = q_4, \quad \text{where} \quad p_j + q_j = 1 \text{ for } j = 1, 2, 3, 4.$$

Using the relations $P\{E_i\} = P\{E_{i-1} E_i\} + P\{\bar{E}_{i-1} E_i\}$,

$$P\{E_{i-1} E_i\} = P\{E_{i-2} E_{i-1} E_i\} + P\{\bar{E}_{i-2} E_{i-1} E_i\}$$

and $$P\{\bar{E}_{i-1} E_i\} = P\{E_{i-2} \bar{E}_{i-1} E_i\} + P\{\bar{E}_{i-2} \bar{E}_{i-1} E_i\},$$

we obtain condition (7), i.e. $$P = PP_1 + QP_2,$$

together with $$P_1 p_1 + Q_1 p_3 = P_1, \tag{8}$$

and $$P_2 p_2 + Q_2 p_4 = P_2. \tag{9}$$

These conditions again ensure that the start of the sequence is a point chosen at random in a longer sequence, and can be written in the form

$$P_1 = \frac{p_3}{1-p_1+p_3}, \quad P_2 = \frac{p_4}{1-p_2+p_4}, \quad P = \frac{P_2}{1-P_1+P_2} = \frac{p_4(1-p_1+p_3)}{p_4(1-p_1+p_3)+(1-p_1)(1-p_2+p_4)}.$$

Thus the problem is reduced to one with four parameters p_1, p_2, p_3, p_4.

To construct the power function under H_2 it is necessary to consider not only the sequences of $2t$ or $2t+1$ groups, but also the subset of such sequences containing a specified number of single E's and of single $\bar{E}$'s. If the r_1 E's are partitioned into t parts of which m_1 parts consist of a single E, then there are $t-m_1$ parts consisting of two or more E's, and the number of times that E follows EE is $r_1-m_1-2(t-m_1)$, i.e. r_1-2t+m_1. Similar results hold if the r_2 $\bar{E}$'s are partitioned into t parts of which m_2 parts consist of single $\bar{E}$'s. Suppose this sequence of $2t$ groups starts with at least two consecutive E's and ends with at least two consecutive $\bar{E}$'s. Then it can easily be seen that the probability of obtaining such a sequence is

$$PP_1q_1q_2p_1^{r_1-2t+m_1}q_4^{r_2-2t+m_2}(p_2q_1)^{t-m_1-1}(p_4q_2)^{t-m_2-1}p_2^{m_2}q_3^{m_1},$$

which we may write for brevity $\dfrac{PP_1}{p_3p_4}z$, where

$$z=p_1^{r_1}q_4^{r_2}\left(\frac{p_1q_3}{p_3q_1}\right)^{m_1}\left(\frac{p_2q_4}{p_4q_2}\right)^{m_2}\left(\frac{p_3p_4q_1q_2}{p_1^2q_4^2}\right)^t,$$

and refer to as sequence (i). The results for sequences (ii) starting EE, ending $E\bar{E}$, (iii) starting $E\bar{E}$, ending $\bar{E}\bar{E}$, (iv) starting $E\bar{E}$, ending $E\bar{E}$ are $\dfrac{PP_1}{p_2p_3}z$, $\dfrac{PQ_1}{p_4q_3}z$, $\dfrac{PQ_1}{p_2q_3}z$ respectively. Four corresponding results are obtained when the E's and $\bar{E}$'s in (i), (ii), (iii) and (iv) are interchanged.

For a sequence of $2t+1$ groups the subdivisions are into (a) $t+1$ groups of r_1 E's containing m_1 single E's and t groups of r_2 $\bar{E}$'s containing m_2 single $\bar{E}$'s, and (b) t groups of r_1 E's containing m_1 single E's and $t+1$ groups of r_2 $\bar{E}$'s containing m_2 single $\bar{E}$'s. Under (a) the sequences considered are those (i) starting EE, ending EE, (ii) starting EE, ending $\bar{E}E$, (iii) starting $E\bar{E}$, ending EE, (iv) starting $E\bar{E}$, ending $\bar{E}E$; and the probabilities of obtaining these sequences are $\dfrac{PP_1}{p_1^2}z$, $\dfrac{PP_1}{p_1^2}\dfrac{q_1}{q_3}z$, $\dfrac{PQ_1}{p_1^2}\dfrac{p_3}{q_3}z$, $\dfrac{PQ_1}{p_1^2}\dfrac{p_3q_1}{q_3^2}z$ respectively. Under (b) four corresponding subdivisions are considered.

Finally, the number of ways in which each such sequence can occur is required. The number of compositions of (r_1-m_1) E's into $(t-m_1)$ parts when no single E occurs is ${}^{r_1-t-1}C_{t-m_1-1}$ for $m_1=0,1,\ldots,t-1$, and is unity for $m_1=t=r_1$. This is the coefficient of $x^{r_1-m_1}$ in $(x^2+x^3+\ldots)^{t-m_1}$. Since the m_1 single E's can occupy the t spaces in ${}^tC_{m_1}$ ways, it follows that the number of compositions of r_1 E's into t parts, m_1 of which are single E's, is ${}^{r_1-t-1}C_{t-m_1-1}\cdot{}^tC_{m_1}$. If this be denoted by $h_1(t,m_1)$, then the number of sequences of $2t$ groups containing m_1 single E's and m_2 single $\bar{E}$'s is $2h_1(t,m_1)h_2(t,m_2)$ and the numbers of sequences of $2t$ groups specified by (i), (ii), (iii), (iv) are given by the product of $h_1(t,m_1)h_2(t,m_2)$ with

$$\frac{(t-m_1)(t-m_2)}{t^2},\quad \frac{(t-m_1)m_2}{t^2},\quad \frac{m_1(t-m_2)}{t^2},\quad \frac{m_1m_2}{t^2}$$

respectively. Similar results hold in the other cases.

Combining the results and summing for all m_1 and m_2, we obtain

$$P\{2t\,|\,r_1r_2H_2\}=\frac{A}{C}\quad\text{and}\quad P\{2t+1\,|\,r_1r_2H_2\}=\frac{B}{C}\quad\text{for}\quad t=1,2,\ldots,r_2\text{ and }r_1\geqslant r_2,$$

where

$$A=\sum_{m_1=0}^{t}\sum_{m_2=0}^{t}\left(\frac{p_1q_3}{p_3q_1}\right)^{m_1}\left(\frac{p_2q_4}{p_4q_2}\right)^{m_2}\left(\frac{p_3p_4q_1q_2}{p_1^2q_4^2}\right)^t\frac{h_1(t,m_1)h_2(t,m_2)}{t^2}\Bigg[(t-m_1)(t-m_2)\left(\frac{PP_1}{p_3p_4}+\frac{QQ_2}{q_1q_2}\right)$$
$$+(t-m_1)m_2\left(\frac{PP_1}{p_2p_3}+\frac{QP_2}{p_2q_1}\right)+m_1(t-m_2)\left(\frac{QQ_2}{q_2q_3}+\frac{PQ_1}{p_4q_3}\right)+m_1m_2\left(\frac{PQ_1+QP_2}{p_2q_3}\right)\Bigg],$$

$$B = \sum_{m_1=0}^{t+1}\sum_{m_2=0}^{t+1}\left(\frac{p_1q_3}{p_3q_1}\right)^{m_1}\left(\frac{p_2q_4}{p_4q_2}\right)^{m_2}\left(\frac{p_3p_4q_1q_2}{p_1^2q_4^2}\right)^t$$
$$\times\left[\frac{h_1(t+1,m_1)\,h_2(t,m_2)}{t(t+1)}\frac{P}{p_1^2}\left\{(t-m_1+1)(t-m_1)P_1+m_1(t-m_1+1)\frac{P_1q_1+Q_1p_3}{q_3}\right.\right.$$
$$\left.+m_1(m_1-1)\frac{Q_1p_3q_1}{q_3^2}\right\}+\frac{h_1(t,m_1)\,h_2(t+1,m_2)}{t(t+1)}\frac{Q}{q_4^2}$$
$$\left.\times\left\{(t-m_2+1)(t-m_2)Q_2+m_2(t-m_2+1)\frac{Q_2p_4+P_2q_2}{p_2}+m_2(m_2-1)\frac{P_2q_2p_4}{p_2^2}\right\}\right]$$

$$C = \sum_{i=1}^{r_s}(A+B), \quad \text{and where} \quad h_i(t,m_i) = {}^{r_i-t-1}C_{t-m_i-1}\,.\,{}^tC_{m_i} \quad \text{for} \quad m_i<t$$
$$= 1 \qquad \text{for} \quad m_i = t,\ i = 1,2.$$

Using conditions (7), (8), (9) these expressions simplify to some extent, and we may note:

(1) If $p_1 = p_2 = p_3 = p_4$, then $P = P_1 = P_2$ and $P\{T \mid r_1r_2H_2\}$ becomes $P\{T \mid r_1r_2H_0\}$, for A reduces to ${}^{r_1-1}C_{t-1}\,{}^{r_2-1}C_{t-1}$ and B to ${}^{r_1-1}C_t\,{}^{r_2-1}C_{t-1}+{}^{r_1-1}C_{t-1}\,{}^{r_2-1}C_t$.

(2) If $p_1 = p_3$ and $p_2 = p_4$, then $P_1 = p_1$ and $P_2 = p_2$ and $P\{T \mid r_1r_2H_2\}$ becomes $P\{T \mid r_1r_2H_1\}$, for A reduces to $\left(\frac{P_2Q_1}{P_1Q_2}\right)^t\left(\frac{P}{P_2}+\frac{Q}{Q_1}\right)^{r_1-1}C_{t-1}\,{}^{r_2-1}C_{t-1}$, and B reduces to

$$\left(\frac{P_2Q_1}{P_1Q_2}\right)^t\left[\frac{P}{P_1}\,{}^{r_1-1}C_t\,{}^{r_2-1}C_{t-1}+\frac{Q}{Q_2}\,{}^{r_1-1}C_{t-1}\,{}^{r_2-1}C_t\right].$$

(3) If $p_1 = p_2$ and $p_3 = p_4$, then $P = P_1 = P_2$ and we have what might be called 'throw-back' dependence, for the result of the ith trial depends only on the result of the $(i-2)$th trial, and in fact the odd and even events in the sequence form independent sequences. The group test will be of little use unless we are interested in both positive and negative dependence as the alternative hypothesis.

(4) If $p_2 = p_3$, then the dependence is of the 'global' type, for each event depends only on the number of 'successes' in the preceding two trials.

6. The distribution of the length of the longest run under hypothesis H_2

This distribution can be obtained in a manner similar to the foregoing, but the formula is even more unwieldy. We shall consider, therefore, only $P\{g \geqslant s \mid r_1r_2H_2\}$. Suppose $h_i(t,m_i\,g<s)$, where $i = 1, 2$, is the number of compositions of r_i elements into t parts of which m_i parts contain one and only one element and such that no part contains s or more than s elements. Then

$$h_i(t,m_i,g<s) = \left[\sum_{j=0}^{t-m_i}(-)^j\,{}^{t-m_i}C_j\,{}^{r_i-t-(s-2)j-1}C_{t-m_i-1}\right]{}^tC_{m_i} \quad \text{for} \quad s = 3,4,\ldots,r_i+1$$
$$= 1 \qquad \text{for} \quad s = 2,$$

where the expression in the square brackets is the coefficient of $x^{r_i-m_i}$ in the expansion of $(x^2+x^3+\ldots+x^{s-1})^{t-m_i}$. Clearly $h(t,m_i,g<r_i+1) = h(t,m_i)$. The number of sequences of $2t$ groups containing m_1 single E's and m_2 single $\bar{E}$'s and such that the longest run of either E's or $\bar{E}$'s has length greater than or equal to s is

$$2[h_1(t,m_1)\,h_2(t,m_2)-h_1(t,m_1,g<s)\,h_2(t,m_2,g<s)],$$

i.e.
$$2\left[\prod_{i=1}^{2}{}^{r_i-t-1}C_{t-m_i-1}\,{}^tC_{m_i}-\prod_{i=1}^{2}\left(\sum_{j=0}^{t-m_i}(-)^j\,{}^{t-m_i}C_j\,{}^{r_i-t-(s-2)j-1}C_{t-m_i-1}\,{}^tC_{m_i}\right)\right].$$

If we denote by A' the expression obtained on replacing $h_1(t, m_1)\, h_2(t, m_2)$ in A (see § 5) by $h_1(t, m_1)\, h_2(t, m_2) - h_1(t, m_1, g < s)\, h_2(t, m_2, g < s)$, and if we obtain B' from B in a similar way, then

$$P\{g \geqslant s \,|\, r_1 r_2 H_2\} = \frac{\sum_{t=1}^{r_1-s+1} (A' + B')}{\sum_{t=1}^{r_2} (A + B)}$$

7. Comparison of certain power curves for the criteria T and g when the alternative hypothesis is double dependence

In plotting the power curves, only the cases where

$$p_1 + p_4 = 1 = p_2 + p_3 \tag{10}$$

have been considered. Using the relations previously obtained, namely

$$P_1 = \frac{p_3}{1 - p_1 + p_3}, \quad P_2 = \frac{p_4}{1 - p_2 + p_4}, \quad P = \frac{P_2}{1 - P_1 + P_2},$$

it follows from (10) that $P_1 = Q_2$, $P_2 = Q_1$ and $P = 0{\cdot}5$. This seems a reasonable case to take for it merely implies symmetry with respect to E and $\bar{E}$. Writing the relations out in full we have $P = 0{\cdot}5$, $P_1 = Q_2$, $P_2 = Q_1$, $p_4 = q_1$, $p_1 = q_4$, $p_2 = q_3$ and $p_3 = q_2$. The power functions can now be expressed in terms of two parameters, say p_1 and p_3. Three sections of this power surface have been considered for $r_1 = r_2 = 10$, namely sections by the planes (*a*) $p_1 = p_3$ (H_1 in Fig. 2), (*b*) $p_3 = 0{\cdot}5$ (H_2' in Fig. 3), (*c*) $4p_3 - 2p_1 = 1$ (H_2'' in Fig. 3).

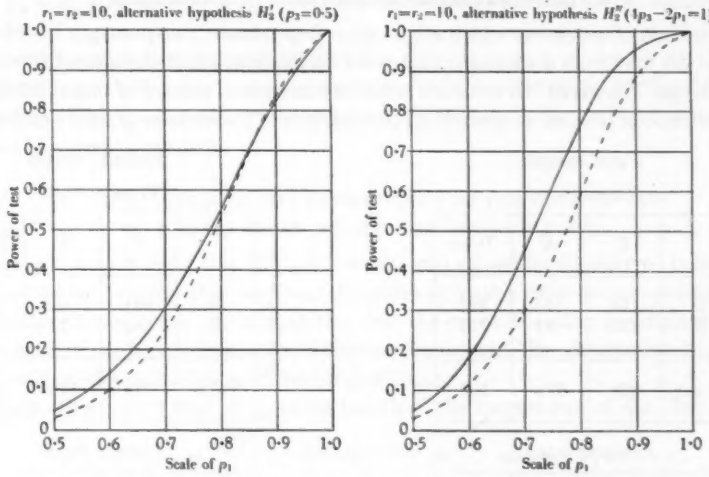


Fig. 3. Case where the alternative hypothesis is double dependence:
— $P\{T \leqslant T_\alpha \,|\, r_1 r_2 H_2\}$ --- $P\{g \geqslant g_\alpha \,|\, r_1 r_2 H_2\}$

(*a*) When $p_1 = p_3$, as it does along the central diagonal in the (p_1, p_3) plane, the dependence in the sequence is of the single kind, and when $p_1 = p_3 = 0{\cdot}5$ then there is independence.

(*b*) When $p_3 = 0{\cdot}5$, then the dependence is of the global kind for the probability of obtaining a success at any trial depends on whether there were 2, 1, 0 successes in the preceding two trials; and in the case taken these probabilities decrease in arithmetic progression, e.g. when $p_1 = 0{\cdot}7$, $p_2 = p_3 = 0{\cdot}5$, $p_4 = 0{\cdot}3$.

(*c*) When $4p_3 - 2p_1 = 1$, the dependence is intermediate between these two types. Typical values of p_1, p_2, p_3, p_4 are shown in the following table:

p_1	0·5	0·6	0·7	0·8	0·9
p_3	0·5	0·55	0·6	0·65	0·7
p_2	0·5	0·45	0·4	0·35	0·3
p_4	0·5	0·4	0·3	0·2	0·1

It will be seen that, apart from the first column in which the events are independent, the probability of an event E occurring decreases according as the results of the two preceding trials are EE, $\bar{E}E$, $E\bar{E}$ or $\bar{E}\bar{E}$.

When the alternative hypothesis is single dependence, then, as we have noted previously, the g criterion is considerably less powerful than the T criterion in detecting departures from randomness. This point is illustrated by the power curves of Fig. 2. The power curves of Fig. 3 show that this is not necessarily the case when the alternative hypothesis is double dependence, for in one of the two cases considered the g criterion would appear to be no less powerful than T in detecting departures from the basic hypothesis.

It is realized that this investigation is not yet complete, but further work is being done on types of dependence likely to occur in practice, and on the possibility of using an estimate of the degree of dependence in a sequence as a test criterion.

8. Illustration of dependence in a sequence

As an illustration of sequences which might show dependence, two passages have been taken at random, the first from a standard text-book on Statistics and the second from Gertrude Stein's writings. The event, E, recorded is the occurrence of a word of one syllable, while $\bar{E}$ is the occurrence of a word of more than one syllable. To eliminate possible end effects the

First sample

Present word

Preceding word	E	$\bar{E}$	Total
E	194	133	327
$\bar{E}$	132	41	173
Total	326	174	500

Present word

Preceding two words	E	$\bar{E}$	Total
EE	108	85	193
$E\bar{E}$	101	32	133
$\bar{E}E$	86	48	134
$\bar{E}\bar{E}$	31	9	40
Total	326	174	500

Second sample

Present word

Preceding word	E	$\bar{E}$	Total
E	301	89	390
$\bar{E}$	89	21	110
Total	390	110	500

Present word

Preceding two words	E	$\bar{E}$	Total
EE	231	72	303
$E\bar{E}$	69	19	88
$\bar{E}E$	70	17	87
$\bar{E}\bar{E}$	20	2	22
Total	390	110	500

first three and the last three words in each sentence have not been counted. The results for the individual sentences have been pooled to obtain the above frequency tables. The total number of words counted was 500 in each case.

Estimates of the probabilities, relative probabilities and degree of dependence (δ) are shown in the following table:

	$\hat{P}$	$\hat{P}_1$	$\hat{P}_2$	$\hat{p}_1$	p_2	$\hat{p}_3$	$\hat{p}_4$	$\hat{\delta}$
1st sample	0·65	0·59	0·76	0·56	0·76	0·64	0·78	−0·17
2nd sample	0·78	0·77	0·81	0·76	0·78	0·80	0·91	−0·04

If we take as our hypothesis that $P_1 = P_2 = P$, then, applying the χ^2-test for independence to the 2×2 tables above, we obtain values for χ^2 of 13·62 for the first sample and 0·49 for the second sample. On this evidence we reject the hypothesis of independence in the first case, but not in the second case. The first sample does, in fact, seem to exhibit a degree of negative dependence of the single kind, i.e. given a monosyllabic (or polysyllabic) word the chance of its being followed by another monosyllabic (or polysyllabic) word is less than it would be if the events were independent. It is not clear whether or not the sequence shows double dependence. Applying the χ^2-test to the 2×4 table we again get a significant value for χ^2, which leads us to reject the hypothesis, $p_1 = p_2 = p_3 = p_4 = P$; but it might still be the case that $p_1 = p_3 = P_1$ and $p_2 = p_4 = P_2$, when the double dependence would reduce to single dependence.

In the second sample it might well be that $P = P_1 = P_2 = p_1 = p_2 = p_3 = p_4$, the fluctuations in the estimates being due to chance, and the χ^2-test applied to the 2×2 and to the 2×4 table confirms this. Thus in Gertrude Stein's work monosyllabic and polysyllabic words would seem to occur at random.

9. Application of the distribution of the longest run to the classical problem of 'runs of luck'

The problem of 'runs of luck' was first put forward and solved by de Moivre. It is concerned with finding the probability that an event E occurs at least s times in succession in a series of r independent trials, when the probability that the event E occurs is constant and equal to p. The problem has been solved using difference equations. The alternative solution given below is based on the distribution of the longest run.

It has been shown in § 2 that, if g_E is the length of the longest run of E's, then

$$P\{g_E \geqslant s \,|\, r_1 r_2\} = \sum_{j=1}^{[r_1/s]} (-)^{j+1} {}_{r_2+1}C_j \, {}_{r-js}C_{r_2} / {}_rC_{r_1}, \quad \text{where} \quad r_1 + r_2 = r.$$

If r (the number of independent trials) is given and $P\{E\}$ is constant and equal to p, then

$$P\{r_1\} = {}_rC_{r_1} p^{r_1} q^{r_2}$$

and

$$\begin{aligned} P\{g_E \geqslant s\} &= \sum_{r_1=s}^{r} P\{g_E \geqslant s \,|\, r_1\} \times P\{r_1\} \\ &= \sum_{r_1=s}^{r} \sum_{j \geqslant 1} (-)^{j+1} {}_{r_2+1}C_j \, {}_{r-js}C_{r_2} \, p^{r_1} q^{r_2} \\ &= \sum_{j \geqslant 1} (-)^{j+1} {}_{r-js}C_{j-1} \, p^{js} q^{j-1} \sum_{r_1 \geqslant js} \frac{r_2+1}{j} \, {}_{r-j(s+1)+1}C_{r_1-js} \, p^{r_1-js} q^{r_2-j+1}. \end{aligned}$$

Using the relation $\sum_{k=0}^{n} {}^{n}C_k p^k q^{n-k} = 1$, we obtain

$$P\{g_E \geqslant s\} = \sum_{j \geqslant 1} (-)^{j+1} {}^{r-js}C_{j-1} p^{js} q^{j-1} \left(1 + \frac{r-j(s+1)+1}{j} q\right)$$

$$= \sum_{j=1}^{[r/s]} (-)^{j+1} \left(p + \frac{r-js+1}{j} q\right) {}^{r-js}C_{j-1} p^{js} q^{j-1}.$$

This can readily be shown to be identical with the solution using difference equations, as given by Uspensky (1937, p. 77).

I wish to thank Dr F. N. David for suggesting the subject and for advice given in the preparation of this paper.

REFERENCES

David, F. N. (1947). *Biometrika*, **34**, 335.
Markoff, A. A. (1913). Supplement to 3rd edition of *Calcul des probabilités*. St Petersbourg.
Mosteller, Frederick (1941). *Ann. Math. Statist.* **12**, 228.
Uspensky, J. V. (1937). *Introduction to Mathematical Probability*. McGraw-Hill Book Company.

SUR LES COURBES DE FRÉQUENCE DE K. PEARSON

PAR M. DUMAS, *Ingénieur en chef de l'Artillerie navale (Marine nationale française)*

1. K. Pearson a indiqué (1895, 1916) comment il était possible, connaissant les rapports homogènes β_1 et β_2 d'une distribution, de calculer l'équation de la courbe des densités d'une loi de probabilité ajustant cette distribution. Il a dû considérer différents cas, suivant la région du plan des $\beta_1\beta_2$ dans laquelle se situe le point intéressant. Ces régions sont indiquées avec précision dans une figure* (K. Pearson, 1916), reproduite en tête de la Part II des *Tables for Statisticians and Biometricians* (K. Pearson, 1931).

Nous nous proposons dans le présent mémoire d'attirer l'attention sur ce qu'il est utile de compléter les indications de la figure susvisée en traçant une nouvelle courbe. Pour l'exposé correspondant, nous conservons les notations utilisées aux p. lx et suivantes de la Part I, édition de 1930, où se trouve un tableau des équations des courbes des fonctions $y(x)$ de K. Pearson. Nous rappelons que l'équation différentielle servant de point de départ à la théorie est

$$\frac{1}{y}\frac{dy}{dx} = \frac{x-a}{c_0+c_1x+c_2x^2}.$$

2. Celles des lignes tracées sur la figure susvisée qui sont les plus intéressantes à considérer du point de vue des formes des courbes représentatives des fonctions $y(x)$ sont:

(a) la droite: $$2\beta_2-3\beta_1-6=0; \tag{1}$$

(b) la cubique et la biquadratique ayant respectivement pour équations:

$$\beta_1(\beta_2+3)^2-4(4\beta_2-3\beta_1)(2\beta_2-3\beta_1-6)=0, \tag{2}$$

$$\beta_1(\beta_2+3)^2(8\beta_2-9\beta_1-12)-4(4\beta_2-3\beta_1)(5\beta_2-6\beta_1-9)^2=0. \tag{3}$$

La biquadratique définie par (3) peut être dite *biquadratique de discontinuité des ordonnées*, car on peut établir ce qui suit: si l'on considère les courbes des fonctions $y(x)$ correspondant respectivement aux β_1 et β_2 de deux points voisins de la biquadratique (3) mais situés de part et d'autre de celle-ci, l'une de ces courbes a une ordonnée nulle en un point ayant pour abscisse une solution de l'équation:

$$c_0+c_1x+c_2x^2=0, \tag{4}$$

tandis que l'autre courbe a, au point correspondant au précédent, une ordonnée infiniment grande.

C'est ainsi par exemple, que la traversée de la branche marquée VIII sur la figure de la Part II susvisée, fait passer la courbe $y(x)$ du type I_U au type I'_J (comme cela est rappelé sur les figures 2 et 4 ci-jointes), en passant d'ailleurs par le type VIII (Fig. 3).

Il est commode, notamment pour la construction des courbes, de remplacer les équations (2) et (3) par des équations paramètriques, en prenant pour paramètres β_1 et le coefficient c_2 de la quadratique (4). On établit en effet que les équations susvisées sont équivalentes à la relation

$$2(1+5c_2)\beta_2=3(1+4c_2)\beta_1+6(1+3c_2), \tag{5}$$

* Cette figure présente notamment l'avantage d'indiquer comme limite de la partie utile du plan des $\beta_1\beta_2$ la droite $\beta_2-\beta_1-1=0$, et d'exclure par suite la droite $4\beta_2-3\beta_1=0$, indiquée dans la figure xxxvi de la Part I des mêmes 'Tables'.

complètée en ce qui concerne la cubique (2) par

$$\beta_1 = -\frac{16c_2(1+3c_2)}{(1+4c_2)^2}; \tag{6}$$

et en ce qui concerne la biquadratique (3) par

$$\beta_1 = \frac{4(1+3c_2)}{(1+4c_2)^2(1+c_2)}. \tag{7}$$

3. L'axe des β_2, la droite $\beta_2-\beta_1-1=0$, la droite (1), et les courbes (2) et (3) limitent différentes régions du plan des $\beta_1\beta_2$; ce sont les seules régions considérées par K. Pearson pour définir ses types de courbes; pour chacune de ces régions, une seule et même expression $y(x)$ est valable. Mais il se trouve que dans certaines de ces régions l'équation correspond à des formes à courbes nettement différentes les unes des autres, car aux points où $y=0$, la pente de la tangente à la courbe $y(x)$ peut être soit *nulle*, soit *infiniment grande*, soit encore, comme état intermédiaire, *finie non nulle*.

Nous avançons que tous les points pour lesquels ces pentes sont finies non nulles, sont ceux situés sur la courbe d'équation

$$\beta_1(\beta_2+3)^2(5\beta_2-6\beta_1-9)-(4\beta_2-3\beta_1)(7\beta_2-9\beta_1-15)^2=0. \tag{8}$$

Soit à démontrer cette proposition.

Rappelons d'abord que la théorie de K. Pearson donne pour les coefficients de la quadratique (4) les expressions

$$c_0 = -\mu_2(1+3c_2) \quad \text{et} \quad c_1 = -\tfrac{1}{2}\epsilon(1+4c_2)\sqrt{(\mu_2\beta_1)} \tag{9}$$

($\epsilon=\pm1$ ayant le signe de μ_3), auxquelles s'ajoute l'équation (5) donnant c_2 en fonction de β_1 et de β_2.

Puis, appelant x_1 et x_2, avec $x_1<x_2$, les solutions réelles de l'équation (4), c'est à dire

$$\left.\begin{matrix} x_1 \\ x_2 \end{matrix}\right\} = \frac{-c_1 \pm \sqrt{(c_1^2-4c_0c_2)}}{2c_2}, \tag{10}$$

posons

$$m_1 = \frac{c_1-x_1}{c_2(x_2-x_1)} \quad \text{et} \quad m_2 = \frac{x_2-c_1}{c_2(x_2-x_1)}; \tag{11}$$

d'après (9) et (10)

$$\left.\begin{matrix} m_1 \\ m_2 \end{matrix}\right\} = \frac{1}{2c_2} \pm \frac{(1+2c_2)(1+4c_2)\sqrt{\beta_1}}{2c_2\sqrt{[\beta_1(1+4c_2)^2+16c_2(1+3c_2)]}}. \tag{12}$$

Considérons maintenant le cas particulier où la solution $y(x)$ a pour expression

$$y(x) = k(x-x_1)^{m_1}(x-x_2)^{m_2};$$

c'est le cas où, pour $\mu_3>0$, la courbe représentative est du type VI de K. Pearson. Dans ce cas

$$\frac{dy(x)}{dx} = \frac{k}{c_2}(x-c_1)(x-x_1)^{m_1-1}(x-x_2)^{m_2-1}.$$

Manifestement, on ne peut avoir à la fois $y(x)$ nul et $\dfrac{dy(x)}{dx}$ fini non nul que pour $x=x_1$ (ou $x=x_2$), à condition d'avoir en même temps $m_1=1$ (ou $m_2=1$), c'est à dire, d'après (12), à condition d'avoir

$$\left(1-\frac{1}{2c_2}\right)^2 = \frac{\beta_1(1+2c_2)^2(1+4c_2)^2}{4c_2^2[\beta_1(1+4c_2)^2+16c_2(1+3c_2)]}.$$

Si dans cette expression on porte la valeur de c_2 déduite de (5), on trouve (8).

Nous ne croyons pas utile d'indiquer ici la discussion complète, seule capable de montrer que la courbe d'équation (8) est valable dans tous les cas et qu'elle est la seule valable.

4. L'équation (8) est l'équation d'une nouvelle biquadratique qui peut être dite, pour la distinguer de celle qui a pour équation (3), *biquadratique de discontinuité des dérivées* en raison de la propriété qui a conduit à la considérer. Ses équations paramètriques sont l'équation (5) complétée par

$$\beta_1 = \frac{2(1+3c_2)(1-2c_2)^2}{(1+4c_2)^2}.$$

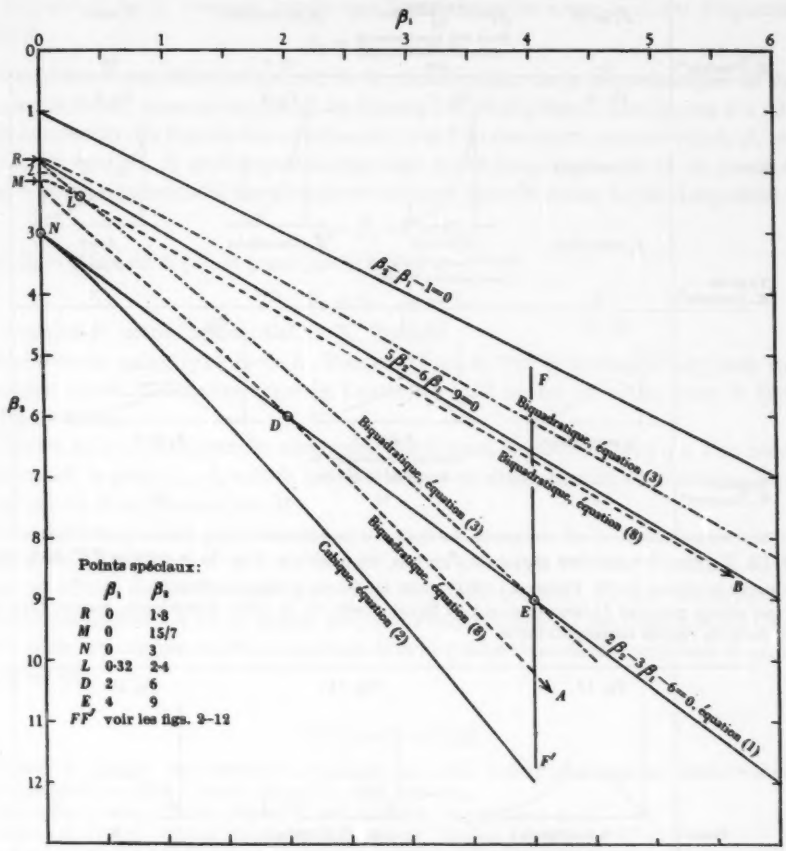


Fig. 1. Limites de régions du plan des $\beta_1\,\beta_2$.

La partie utile de cette biquadratique a deux branches dont l'allure générale est celle des deux branches de la biquadratique (3); l'une d'elles est asymptote à la droite d'équation:

$$5\beta_2 - 6\beta_1 - 9 = 0;$$

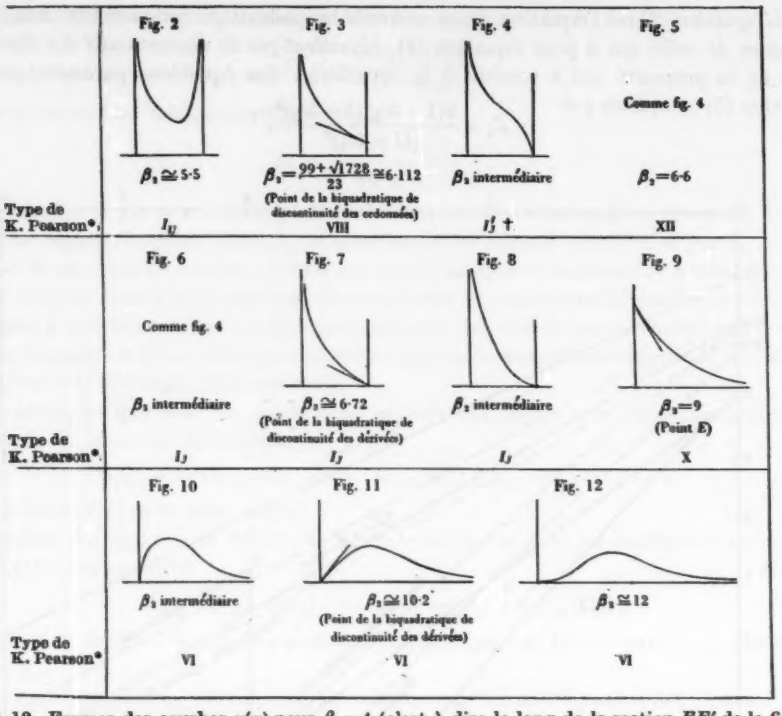
l'autre est pour β_2 infiniment grand, asymptote à

$$\beta_1 = 39{\cdot}2;$$

cette dernière équation est à comparer aux suivantes

$\beta_1 = 32$, qui est celle de l'asymptote à la cubique (2);

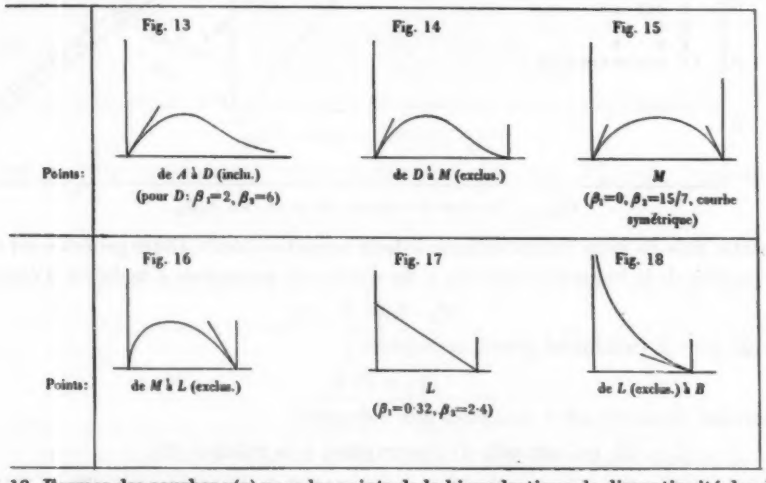
$\beta_1 = 50$, qui est celle d'une asymptote à la biquadratique (3).



Figs. 2–12. Formes des courbes $y(x)$ pour $\beta_1 = 4$ (c'est à dire le long de la section FF' de la figure 1).

*** D'après la figure de K. Pearson (1931) visée au § 1 du présent mémoire.**

† Nous avons marqué I_J' bien que, sur la figure visée en la note précédente, aucun type ne soit indiqué dans la région correspondante.



Figs. 13–18. Formes des courbes $y(x)$ pour les points de la biquadratique de discontinuité des dérivées.

Comme on le voit sur la figure 1, ces deux branches se raccordent tangentiellement à l'axe des β_2 au point M

$$\beta_1 = 0, \quad \beta_2 = \tfrac{15}{7};$$

l'une d'elles coupe la biquadratique de discontinuité des ordonnées au point

$$\beta_1 = 0{\cdot}32, \quad \beta_2 = 2{\cdot}4,$$

qui est le point L de K. Pearson, tandis que l'autre branche coupe la droite d'équation (1) au point D

$$\beta_1 = 2, \quad \beta_2 = 6.$$

5. Pour donner une idée de l'intérêt de la considération de la biquadratique de discontinuité des dérivées, nous avons établi les figures 2 à 18 ci-jointes. Les figures 2 à 12 indiquent la succession des formes des courbes $y(x)$ que l'on rencontre dans le cas où, β_1 restant constamment égal à 4, β_2 croît depuis les environs de 5·5 jusqu'à ceux de 12; en raison de ces valeurs, le point représentatif reste toujours compris entre le droite limite d'équation

$$\beta_2 - \beta_1 - 1 = 0,$$

et la cubique d'équation (2), et passe par le point

$$\beta_1 = 4, \quad \beta_2 = 9,$$

qui est le point E, point exponentiel, de K. Pearson.

On remarquera qu'au type I_J de K. Pearson (Figs. 6, 7 et 8) correspondent trois courbes d'allures nettement différentes l'une de l'autre, et qu'il en est de même pour le type VI, (Figs. 10, 11 et 12.)

Les figures 13 à 18 indiquent la succession des formes de courbes $y(x)$ que l'on rencontre dans le cas où le point β_1, β_2 suit la biquadratique de discontinuité des dérivées depuis A (Fig. 1) jusqu'à B en passant par M.

Nous comptons publier prochainement un exposé complet de la théorie mathématique conduisant aux fonctions $y(x)$; à cet exposé seront joints un tableau d'environ 34 figures donnant les allures des courbes $y(x)$ pour tous les points possibles du plan des $\beta_1\beta_2$ avec, en particulier, les figures 2 à 18, et toutes précisions analytiques sur les fonctions $y(x)$, notamment sur celles qui correspondent aux points de la biquadratique de discontinuité des dérivées (Figs. 13 à 18).

BIBLIOGRAPHIE

1. PEARSON, K. (1895). Mathematical contributions to the theory of evolution. Skew variation in homogeneous material. *Philos. Trans.* A, **186**, 343–414.
2. PEARSON, K. (1901). *Philos. Trans.* A, **197**, 443–59. Supplément de 1.
3. PEARSON, K. (1916). *Philos. Trans.* A, **216**, 429–57. Second supplément 1.
4. PEARSON, K. (1930). *Tables for Statisticians and Biometricians.* Part I (Edition de 1930).
5. PEARSON, K. (1931). Part II de 4.
6. ELDERTON, W. PALIN (1937). *Frequency-Curves and Correlation.* Cambridge University Press.

THE DISTRIBUTION OF THE EXTREME DEVIATE FROM THE SAMPLE MEAN AND ITS STUDENTIZED FORM*

By K. R. NAIR

1. Introduction

Denote by $x_{(1)} \dots x_{(n)}$ a random sample of n observations drawn from any statistical universe. Let $x_1, \dots, x_n$ be the same sample arranged in ascending order of magnitude so that x_r is the rth ranked (or ordered) variate in the sample $\{x_{(i)}\}$.

Various authors have studied the sampling distribution of x_r. It takes a simple form when $r = 1$ or n. Thus, if $f(x)$ is the probability function of the universe from which $x_{(i)}$'s are drawn, the probability that $x_n \leqslant X$ is given by

$$P(X) = \left[\int_{-\infty}^{x} f(x)\,dx\right]^n. \tag{1}$$

If the parent universe is normal with mean μ and standard deviation σ

$$f(x) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \tag{2}$$

The probability that $x_n \leqslant X$ is the same as that of

$$v_n = \frac{x_n - \mu}{\sigma} \leqslant \frac{X-\mu}{\sigma} = V \tag{3}$$

and may be written $$P(X) = P(V) = \left(\frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{V} e^{-\frac{1}{2}x^2}dx\right)^n. \tag{4}$$

Tippett (1925) tabulated values of $P(V)$ for n ranging between 3 and 1000. E. S. Pearson supplemented this by a table of percentage points of V, published as Table XXI *bis* in *Tables for Statisticians and Biometricians*, Part II, to facilitate tests of significance of a single outlying observation x_1 or x_n, when both μ and σ are known.

Very often we do not know either μ or σ or both. When μ alone is unknown, Irwin (1925) suggested a test for a single outlier x_n based on the statistic $(x_n - x_{n-1})/\sigma$. If there are k large outliers, his test would involve $(x_{n-k+1} - x_{n-k})/\sigma$ and, for k small outliers, the statistic $(x_{k+1} - x_k)/\sigma$.

Another statistic commonly used to test a single outlier, when μ is unknown, is one based on the range, viz. $(x_n - x_1)/\sigma$ (see 'Student', 1927).

Intuitively, one feels that a better criterion than $(x_n - x_{n-1})/\sigma$ or $(x_n - x_1)/\sigma$ when there is only *one* outlier x_n will be the extreme deviate $(x_n - \bar{x})/\sigma$ proposed by McKay (1935). The range $(x_n - x_1)$ will certainly be a surer guide than $(x_n - x_{n-1})$ or $(x_n - \bar{x})$ if *both* x_1 and x_n are outliers.

In general, if there are k outliers at the lower end and l outliers at the upper end, a suitable criterion will be that based on the difference of the means of the first k and the last l outliers. When the outliers are all at one end, say the first k, we should write $k + l = n$. The distribution of this general statistic and some of its special cases are considered in § 5.

* Part of a thesis approved for the degree of Ph.D. of the University of London.

All these test criteria should be handled with extreme caution in practice. It may be appropriate here to quote Pearson & Chandrasekar's (1936) warning:

> To base the choice of the test of a statistical hypothesis upon an inspection of the observations is a dangerous practice; a study of the configuration of a sample is almost certain to reveal some feature, or features, which are exceptional if the hypothesis is true.... By choosing the feature most unfavourable to the hypothesis out of a very large number of features examined, it will usually be possible to find some reason for rejecting the hypothesis.

Let us now consider the case where both μ and σ are unknown. Their sample estimates are

$$\bar{x} = \sum_1^n (x_i)/n \quad \text{and} \quad s = \sqrt{\left[\sum_1^n (x_i - \bar{x})^2/(n-1)\right]}. \tag{5}$$

If x_n is to be tested as an outlier, the obvious course is to consider the modified McKay criterion $(x_n - \bar{x})/s$. The sampling distribution of this test criterion is not known, but Thompson (1935) suggested an alternative method based on the sampling distribution of

$$\tau_{(i)} = (x_{(i)} - \bar{x})/s, \tag{6}$$

where (see above) $x_{(i)}$ is a *random* member of the sample. He showed that the distribution of τ can be derived from that of t (with $(n-2)$ degrees of freedom) using the relation

$$\tau = \frac{(n-1)\,t}{\sqrt{[nt^2 + n(n-2)]}}. \tag{7}$$

On a critical examination of Thompson's criterion Pearson & Chandrasekar found that it is effective only when there is essentially *one* outlier and not many as Thompson seemed to believe.

Suppose now that μ is known and σ is unknown. Thompson's criterion could be extended to this situation if we calculate

$$\tau'_{(i)} = (x_{(i)} - \mu)/s', \tag{8}$$

where $s' = \sqrt{\left[\sum_1^n (x_i - \mu)^2/n\right]}$. The distribution of τ' can be obtained from that of t (with $(n-1)$ degrees of freedom) using the relation

$$\tau' = \frac{t\sqrt{n}}{\sqrt{(t^2 + n - 1)}}. \tag{9}$$

We now come to the main theme of this paper which is to consider a method of studentization different from Thompson's. If an estimate s_ν of the unknown σ is available with ν degrees of freedom *independent* of the sample $\{x_{(i)}\}$, a test for a single outlier can be made using $(x_n - \bar{x})/s_\nu$ if μ is unknown, and using $(x_n - \mu)/s_\nu$ if μ is known. The distributions of these studentized test criteria can be derived by Hartley's (1944) method.

To test whether there are two outliers, one at each end, the studentized range $(x_n - x_1)/s_\nu$ will be a suitable criterion. Pearson & Hartley (1943) have prepared the necessary tables for this test.

In this paper attention is mainly concentrated on the McKay statistic, $u = (x_n - \bar{x})/\sigma$ (or $(\bar{x} - x_1)/\sigma$) and its studentized form $(x_n - \bar{x})/s_\nu$ (or $(\bar{x} - x_1)/s_\nu$). It is shown that the distribution of u can be obtained by a more direct method than was employed by McKay and that it can be reduced to certain integrals which have recently been termed G-functions by Godwin (1945) in his representation of the distribution of the mean deviation. Tables of the probability integral of u and of its studentized form have been prepared.

Apart from serving as a criterion for rejection of an outlying observation in a 'normal' sample when μ and σ are unknown, the studentized form of u has useful application in judging the significance of a single outstanding treatment (best or worst) in a group of treatments tried out in a designed experiment. Some illustrations are given in the paper.

2. Distribution of the extreme deviate

Assuming, without loss of generality, that $\mu = 0$ and $\sigma = 1$ in the normal probability function (2), the joint distribution of the ordered variates $x_1, \ldots, x_n$ is

$$\frac{n!}{[\sqrt{(2\pi)}]^n} \exp\left[-\frac{1}{2}\sum_1^n x_i^2\right] \prod_1^n dx_i. \tag{10}$$

Making an orthogonal transformation of $x_1, \ldots, x_n$ into $y_1, \ldots, y_n$ defined by

$$\left.\begin{aligned} y_i &= \frac{1}{\sqrt{[i(i+1)]}}\{ix_{i+1} - (x_i + \ldots + x_1)\} \quad (i = 1, \ldots, (n-1)) \\ &= \frac{z_i}{\sqrt{[i(i+1)]}} \quad \text{(say)}, \\ y_n &= \frac{1}{\sqrt{n}}\sum_1^n (x_i) = \sqrt{n}\,\bar{x}, \end{aligned}\right\} \tag{11}$$

it follows that

$$\sum_1^n x_i^2 = \sum_1^n y_i^2 = \sum_1^{(n-1)} \frac{z_i^2}{i(i+1)} + n\bar{x}^2, \tag{12}$$

$$\prod_1^n dx_i = \prod_1^n dy_i = \frac{1}{(n-1)!} d\bar{x} \prod_1^{(n-1)} dz_i. \tag{13}$$

The joint distribution of $\bar{x}, z_1, \ldots, z_{(n-1)}$ is therefore

$$\frac{n}{[\sqrt{(2\pi)}]^n} \exp\left[-\frac{n\bar{x}^2}{2} - \frac{1}{2}\sum_1^{(n-1)} \frac{z_i^2}{i(i+1)}\right] d\bar{x} \prod_1^{(n-1)} dz_i. \tag{14}$$

Integrating out for $\bar{x}$ from $-\infty$ to $+\infty$, the joint distribution of $z_1, \ldots, z_{(n-1)}$ is

$$\frac{\sqrt{n}}{[\sqrt{(2\pi)}]^{n-1}} \exp\left[-\frac{1}{2}\sum_1^{(n-1)} \frac{z_i^2}{i(i+1)}\right] \prod_1^{(n-1)} dz_i. \tag{15}$$

It will be noticed that

$$z_{(n-1)} = n(x_n - \bar{x}) = nu. \tag{16}$$

To derive the distribution of $z_{(n-1)}$ or of u, we have to integrate out for $z_1, \ldots, z_{(n-2)}$ in (15). The ranges of the z_i are from 0 to $+\infty$ but they are interlinked by the inequalities

$$0 \leqslant z_1 \leqslant \ldots \leqslant z_{(n-2)} \leqslant z_{(n-1)} = nu. \tag{17}$$

At this stage we bring in the G-functions defined by Godwin as follows:

$$G_0(x) = 1, \quad G_r(x) = \int_0^x \exp\left[-\frac{t^2}{2r(r+1)}\right] G_{r-1}(t)\,dt. \tag{18}$$

Integrating out $z_1, \ldots, z_{n-2}$ from (15) between the limits (17), the distribution of u can be written

$$f_n(u) = \frac{n\sqrt{n}}{[\sqrt{(2\pi)}]^{n-1}} \exp\left[-\frac{nu^2}{2(n-1)}\right] G_{n-2}(nu), \tag{19}$$

and the probability integral of u is

$$P_n(u) = \int_0^u f_n(u)\,du = \frac{\sqrt{n}}{[\sqrt{(2\pi)}]^{n-1}} G_{n-1}(nu). \tag{20}$$

Proceeding by an entirely different method, McKay derived an expression for $P_n(u)$ in the form

$$P_n(u) = \exp\left[-\frac{1}{2n}\frac{d^2}{du^2}\right]\left(\frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^u e^{-\frac{1}{2}x^2}\,dx\right)^n, \tag{21}$$

and obtained a recurrence formula

$$f_n(u) = P_n'(u) = \frac{n}{\sqrt{(2\pi)}} \exp\left[-\frac{nu^2}{2(n-1)}\right] \sqrt{\frac{n}{(n-1)}}\, P_{n-1}\left(\frac{nu}{n-1}\right). \tag{22}$$

The integral-power appearing on the right-hand side of (21) is the probability integral of $(x_n-\mu)/\sigma$ given in (4), revealing an interesting connexion between the probability integrals of $(x_n-\bar{x})/\sigma$ and $(x_n-\mu)/\sigma$.

McKay did not attempt to tabulate values of $P_n(u)$, but for getting the upper 5 or 1 % level of u, that is, when $P_n(u) = 0{\cdot}95$ or $0{\cdot}99$, suggested the approximation

$$P_n(u) \sim 1 - \frac{n}{\sqrt{(2\pi)}} \int_{u\sqrt{[n/(n-1)]}}^{\infty} e^{-\frac{1}{2}t^2}\, dt. \tag{23}$$

With the help of exact values of $P_n(u)$ tabulated in the next section, it has been possible to examine the closeness of this approximation. The agreement is remarkably good.

He also gave an approximate formula when u is very small, viz.

$$P_n(u) \sim \frac{n}{2}\, n! \left(\frac{u}{\sqrt{(2\pi)}}\right)^{n-1} \quad (n>2). \tag{24}$$

This does not give a good agreement with the exact values.

A better approximation was found to be

$$P_n(u) \sim \frac{\sqrt{n}}{(n-1)!} \left(\frac{nu}{\sqrt{(2\pi)}}\right)^{n-1} \quad (n \geqslant 2), \tag{25}$$

which could further be improved to

$$P_n(u) \sim \frac{\sqrt{n}}{(n-1)!} \left(\frac{nu}{\sqrt{(2\pi)}}\right)^{n-1} \left\{1 - \frac{n(n-1)\,u^2}{2(n+1)}\right\} \quad (n \geqslant 2). \tag{26}$$

A comparison of these three approximations with the exact values has been made for a few small values of u and with different values of n in Table 4.

From (25) and (26) we can derive the following approximation for $G_r(x)$ when x is small:

$$G_r(x) \sim \frac{x^r}{r!}, \tag{27}$$

$$G_r(x) \sim \frac{x^r}{r!} \left\{1 - \frac{rx^2}{2(r+1)\,(r+2)}\right\}. \tag{28}$$

3. Tables of the probability integral of the extreme deviate

The starting-point was manuscript tables of the G_r-functions ($r = 2, \ldots, 8$) prepared by Godwin (1945) and Hartley (1945) in the course of their work on the mean deviation. The functions they had actually tabulated were multiples of the G-functions, namely, $G^*_r(x) = b_r G_r(x)$. Keeping $n = r+1$, the probability integral of u given in (20) may be written

$$P_{r+1}(u) = c_r G^*_r[(r+1)\,u]. \tag{29}$$

Values of b_r and c_r are given below.

r	b_r	c_r
2	$480\,\pi^{-1}\,2^{-4}$	25515518×10^{-16}
3	$480^2\pi^{-4}\,2^{-4}$	61380797×10^{-13}
4	$480^3\pi^{-3}\,2^{-4}$	14297045×10^{-15}
5	$480^4\pi^{-4}\,2^{-3}$	65256785×10^{-18}
6	$480^5\pi^{-3}\,2^{-4}$	29368910×10^{-20}
7	$480^6\pi^{-4}\,2^{-6}$	13081952×10^{-22}
8	$480^7\pi^{-4}\,2^{-7}$	57814607×10^{-25}

The task was therefore to convert the G^*-tables from regular arguments in x to regular arguments in u. In order to avoid excessive interpolation for $x = (r+1)u$ in the G_r^*-tables it was decided to adopt a pivotal interval of 0·05 for u, which required selecting for x the pivotal intervals 0·15, 0·20, 0·25, 0·30, 0·35, 0·40 and 0·45 for $r = 2, 3, \ldots, 8$ respectively. The x-interval in Godwin & Hartley's tables was 0·05 for $r = 2, 3, 4$; and 0·10 for $r = 5, 6, 7, 8$. At the intervals 0·35 and 0·45 for x in G_6 and G_8, the half-way point formula of Lagrangian interpolation was applied on the 0·1 interval values of $G_6^*(x)$ and $G_8^*(x)$. Values of $P_n(u)$ for $n = 3$ to 9 were then obtained at intervals of 0·05 and then subtabulated on the National Accounting Machine at intervals of 0·01. These values correct to six decimal places are given in Table 1 at the end of the paper. All the calculations were undertaken by the National Physical Laboratory, Mathematical Division, under the direction of Dr Goodwin and Mr Vickers.

Table 2, printed after Table 1 at the end of the paper, gives percentage points of the extreme deviate, at twelve different levels. These were calculated by interpolation from Table 1.

We are now in a position to examine the adequacy of the approximations (23)–(26). Taking (23) first, the following Table 3 shows the exact and approximate values of the upper 5 and 1 % points.

Table 3. *Comparison of exact and approximate upper percentage points of u*

n	5 %		1 %	
	Exact	Approx. (23)	Exact	Approx. (23)
3	1·7375	1·7376	2·2152	2·2152
4	1·9409	1·9413	2·4310	2·4310
5	2·0801	2·0807	2·5743	2·5743
6	2·1843	2·1854	2·6794	2·6795
7	2·2667	2·2683	2·7613	2·7615
8	2·3344	2·3364	2·8279	2·8281
9	2·3916	2·3940	2·8837	2·8839

The agreement is very good. This high accuracy of his formula was apparently not known to McKay and has now been established by computation of the exact results. It may enable us to extend the range of the present Table 2 for obtaining upper percentage points of u for sample sizes beyond 9.

Approximations (24), (25) and (26) were compared with the exact value of $P_n(u)$ for $u = 0{\cdot}05$, 0·10 and 0·20 and $n = 3, 4, 5$. The results are given in Table 4 where (a), (b), (c) stand for the approximations (24), (25), (26) respectively and (d) stands for the exact value. (a) and (b) are identical when $n = 3$, but it will be seen that when n exceeds 3, (b) gives much closer agreement than (a).

With the help of the approximations (23) and (26) it may also be possible to extend Godwin & Hartley's tables of upper and lower percentage points of the mean deviation for sample size beyond 10.

To illustrate the use of Table 1 or 2 in testing the significance of an outlying observation an example is given below.

Example. This is taken from McKay's paper. In the course of routine testing of a standard leather product of a tannery, five parallel tests yielded the following values for the hide substance content of the leather specimens:

32·44, 36·45, 39·64, 40·13, 41·09.

The first observation appears unduly low.

Long experience of the product in question has established a value of 2·2 for the standard deviation σ. We get $u = (37{\cdot}95 - 32{\cdot}44)/2{\cdot}2 = 2{\cdot}50$.

Value of P_5 (2·50) found from Table 1 is 0·987 031. The probability of getting a value 2·50 or larger for u is thus about 0·013. If we use Table 2 only, we see that this probability is between 0·025 and 0·01. On the 5 % level of significance, we conclude that the observation 32·44 is anomalous.

Table 4. *Comparison of approximate and exact values of $P_n(u)$ when u is small*

n	Method	$u = 0{\cdot}05$	$u = 0{\cdot}10$	$u = 0{\cdot}20$
3	(*a*)	0·003 101	0·012 405	0·049 620
	(*b*)	0·003 101	0·012 405	0·049 620
	(*c*)	0·003 095	0·012 312	0·048 131
	(*d*)	0·003 095	0·012 312	0·048 166
4	(*a*)	0·000 127	0·001 016	0·008 127
	(*b*)	0·000 169	0·001 355	0·010 836
	(*c*)	0·000 169	0·001 338	0·010 316
	(*d*)	0·000 169	0·001 338	0·010 334
5	(*a*)	0·000 004	0·000 071	0·001 133
	(*b*)	0·000 009	0·000 148	0·002 360
	(*c*)	0·000 009	0·000 145	0·002 202
	(*d*)	0·000 009	0·000 145	0·002 210

(*a*) McKay's approximation (24).
(*b*) New approximation (25).
(*c*) New approximation (26).
(*d*) Exact value.

4. THE STUDENTIZED INTEGRAL OF THE EXTREME DEVIATE

We have seen that in order to apply the test for the extreme deviate $x_n - \bar{x}$, it is necessary to know the standard deviation σ of the parent normal population. We shall now assume that an estimate s_ν of the unknown σ, equal to the square-root of a variance based on ν degrees of freedom and *independent* of the sample $(x_1, \ldots, x_n)$ is available.

Let ${}_\nu P_n(Q)$ denote the probability that $u_\nu = \dfrac{x_n - \bar{x}}{s_\nu} \leqslant Q$. When $\nu \to \infty$ this tends to the probability $P_n(Q)$ that $u = \dfrac{x_n - \bar{x}}{\sigma} \leqslant Q$ given by (20).

Using Hartley's expansion for studentized integrals up to terms in ν^{-2} we may write

$$ {}_\nu P_n(Q) = a_0 + a_1/\nu + a_2/\nu^2, \tag{30}$$

where $a_0 = P_n(Q)$,

and
$$a_1 = \tfrac{1}{4}\{Q^3 P_n''(Q) - Q P_n'(Q)\}, \tag{31}$$

$$a_2 = \tfrac{1}{96}\{3Q^4 P_n^{\mathrm{iv}}(Q) - 2Q^3 P_n'''(Q)\} - \tfrac{1}{8} a_1. \tag{32}$$

Godwin & Hartley had prepared manuscript tables of $G_r^{*\prime}(x)$, the first derivative of $G_r^*(x)$. Values of $P'_{r+1}(Q)$ could be calculated with the help of these tables, using the formula

$$P'_{r+1}(Q) = \frac{24}{W}(r+1)\,c_r\,G_r^{*\prime}(x), \tag{33}$$

where $x = (r+1)\,Q = nQ$ and W is the interval of x in $G_r^{*\prime}(x)$.

Tables of $P'_n(Q)$ for $n = 3, \ldots, 9$ were prepared at the pivotal interval 0·05 for Q just in the same way as $P_n(u)$ was calculated from $G_r^*(x)$.

Values of the 2nd, 3rd and 4th derivatives of $P_n(Q)$ were obtained from $P'_n(Q)$ by numerical differentiation on the National Accounting Machine. Substituting these in (31) and (32), values of a_1 and a_2 were obtained at a pivotal interval of 0·05. Table 5, printed at the end of the paper, gives the values of a_0, a_1 and a_2 at 0·20 interval for the argument Q. Values of ${}_\nu P_n(Q)$ can easily be read off from this table using formula (30). For sake of convenience in making tests of significance, 5 and 1 % points have been calculated in Table 6, which is also reproduced at the end of the paper.

Since numerical differentiation had to be used in calculating even a_1 and a_2, it was not feasible to bring in the next higher terms $a_3\nu^{-3}$ and $a_4\nu^{-4}$ for calculating ${}_\nu P_n(Q)$ as that would have involved calculation of the 5th to 8th derivatives of $P_n(Q)$ by numerical differentiation.

Values of a_1 are given in Table 5 to four decimals for all values of n except $n = 3$, for which five decimals have been supplied. The maximum error is not more than 2 in the last place.

Values of a_2 are given to three decimals for all values of n except $n = 9$, for which only two decimals are retained, when Q exceeds 1·00. The maximum error is not more than 4 in the last place.

This degree of error is inherent in the method of numerical differentiation which was used, but will not seriously affect values of ${}_\nu P_n(Q)$ when ν is moderately large. It is essential, however, to check against systematic errors and for this purpose the following checks proved quite useful:

(i) When $n = 3$, it is comparatively easy to calculate the exact values of a_1 and a_2 directly from the expressions

$$a_1 = \frac{3\sqrt{3}}{8\pi}\,Q\left\{3Qe^{-3Q^2} - (3Q^2+2)\,e^{-\frac{1}{2}Q^2}\int_0^{\frac{1}{2}Q} e^{-t^2}\,dt\right\},$$

$$a_2 = \frac{3\sqrt{3}}{256\pi}\,Q^3\left\{3Q(189Q^2-16)\,e^{-3Q^2} - (27Q^4-42Q^2-8)\,e^{-\frac{1}{2}Q^2}\int_0^{\frac{1}{2}Q} e^{-t^2}\,dt\right\} - \tfrac{1}{8}a_1.$$

In the table below a small panel of values has been calculated from the above exact formulae and compared with the values given in Table 5. The agreement is very satisfactory.

Table 7

Q	a_1		a_2	
	Exact	Approximate	Exact	Approximate
1	−0·38720	−0·38720	+0·275	+0·275
2	−0·25540	−0·25540	−0·550	−0·551
3	−0·01866	−0·01865	−0·324	−0·323
4	−0·00022	−0·00021	−0·014	−0·014

(ii) It is easy to see by partial integration that the expected value of $(x_n - \bar{x})/\sigma$ or $(\bar{x} - x_1)/\sigma$ can be expressed in terms of integrals of a_1 and a_2 over Q from 0 to ∞. Thus,

$$E\left(\frac{x_n - \bar{x}}{\sigma}\right) = \int_0^\infty Q P'_n(Q)\, dQ$$

$$= -\frac{4}{3}\int_0^\infty a_1 dQ$$

$$= -\frac{32}{25}\int_0^\infty a_2 dQ,$$

where $P'_n(Q)$ is given by (33).

Since the expected values of $x_n - \bar{x}$ and $\bar{x} - x_1$ are equal, the expected value of Range/σ or $(x_n - x_1)/\sigma$ is twice that of the extreme deviate. Hence

$$\frac{\text{Mean range}}{\sigma} = -\frac{8}{3}\int_0^\infty a_1 dQ$$

$$= -\frac{64}{25}\int_0^\infty a_2 dQ.$$

By Gregory's formula for numerical quadrature we can obtain values of the integrals $\int a_1 dQ$, $\int a_2 dQ$ by simple summation. The value of the mean range for $n = 3$ to 9 thus obtained should agree reasonably with the exact values given in Table XXII of *Tables for Statisticians and Biometricians*, Part II. Table 8 below presents the values of the mean range obtained from a_1 and from a_2, alongside the exact values.

Table 8. *Values of mean range* ($\sigma = 1$)

n	Using a_1	Using a_2	Exact value
3	1·69255	1·691	1·69257
4	2·0552	2·053	2·05875
5	2·3236	2·318	2·32593
6	2·5361	2·518	2·53441
7	2·7081	2·687	2·70436
8	2·8504	2·862	2·84720
9	2·9705	2·99	2·97003

Only as many decimals have been retained in columns (2) and (3) of Table 8 as were available for the values of a_1 and a_2 in Table 5. The agreement is satisfactory for our purpose.

Two examples of the use of studentized probability integral of the extreme deviate are given below.

Example 1. This is taken from Snedecor's *Statistical Methods* (1946, 4th ed., p. 266).

A randomized block experiment with four strains of wheat A, B, C, D and five replications gave the following mean yield in pounds per plot.

A	B	C	D
34·4	34·8	33·7	28·4

An interesting feature is the similarity of the first three means. Analysis of variance showed significant differences among the four means, as Table 9 will show.

The variance ratio F for strains against error is $44{\cdot}82/2{\cdot}19 = 20{\cdot}5$ which is much larger than $F_{0{\cdot}001} = 10{\cdot}8$. As Snedecor says: 'one suspects that the highly significant F is attributable largely to the small yield of D'. But according to him 'no definite probability statements can be made about contrasts suggested by the data'. What he means perhaps is that a test of significance of the difference between 28·4 and the mean of 34·4, 34·8 and 33·7 by the usual t-test

$$t = \frac{34{\cdot}3 - 28{\cdot}4}{\sqrt{[2{\cdot}19(\frac{1}{15} + \frac{1}{5})]}} = 7{\cdot}7,$$

with 12 degrees of freedom is not valid, as 28·4 is the smallest mean. The appropriate criterion in this situation is the studentized extreme deviate which in this case is

$$\frac{\bar{x} - x_1}{s_\nu} = \frac{32{\cdot}8 - 28{\cdot}4}{\sqrt{(\frac{1}{5} \times 2{\cdot}19)}} = 6{\cdot}7,$$

with $n = 4$ and $\nu = 12$. Referring to Table 5 (p. 142 below) we find that $Q = 6{\cdot}7$ is far beyond the limits of the table showing that it is a very exceptional value.

Table 9. *Analysis of variance*

Source of variation	Degrees of freedom	Sum of squares	Variance
Blocks	4	21·46	5·36
Strains	3	134·45	44·82
Error	12	26·26	2·19
Total	19	182·17	—

Having thus concluded that the smallest mean 28·4 is significantly smaller than the other three means, we are justified in saying that D is definitely inferior to A, B and C. Of course, as a general rule, no categorical conclusions can be drawn from the results of a single experiment and only by repeating the experiment a number of times and seeing whether D turns out to give the smallest mean every time can our conclusion from the first results be definitely established.

Example 2. This is artificially built up from Example 1 by changing the error variance from 2·19 to 13·00. The latter gives a standard error per plot of 11 % which is not too excessive a value to expect in ordinary field experiments. The variance ratio for strains against this error is $F = 44{\cdot}82/13{\cdot}00 = 3{\cdot}45$ which is not significant at the 5 % level. The conclusion is therefore that there are no significant differences among the means of the four strains A, B, C, D.

But if we compare the smallest mean against the general mean and calculate the studentized extreme deviate

$$\frac{\bar{x} - x_1}{s_\nu} = \frac{32{\cdot}8 - 28{\cdot}4}{\sqrt{(\frac{1}{5} \times 13{\cdot}00)}} = 2{\cdot}7,$$

we find from Table 6 (see p. 143 below) that the probability of getting this or a larger value when $n = 4, \nu = 12$ lies between 0·05 and 0·01. On the 5 % level, therefore, 28·4 is significantly smaller than the general mean, indicating that D is inferior to A, B and C. Although this is an artificially constructed example, it helps to illustrate a situation which may occasionally arise.

5. Use of G-functions in the distributions of a class of statistics based on ordered variates

Godwin introduced the G-functions primarily to obtain the distribution of the mean deviation from the mean. We have seen its use in the distribution of the extreme deviate from the mean. We shall now consider certain other statistics based on the ordered variates $x_1, \ldots, x_n$ in whose sampling distributions the G-function appears. The distributions derived are in the form of single or double integrals involving the G-functions in the integrand. Only by numerical integration using the G-function tables can the probability integrals of these distributions be calculated. In many cases where only upper and lower percentage points are required, the approximations (27) and (28) to the G-functions may be quite adequate.

Let us consider a general statistic

$$\delta = (x_n + \ldots + x_{n-l+1})/l - (x_1 + \ldots + x_k)/k, \tag{34}$$

where $x_1 \leqslant \ldots \leqslant x_n$ and $k + l \leqslant n$.

When $k = l$, δ is same as the statistic suggested by Jones (1946) for a quick estimation of σ in certain situations where (say) only 5 % of the observations at the two ends are available. As special cases of this statistic we have the range and the mean deviation from the median (see, for example, Nair, 1947).

When $k + l = n$, δ may be written in three different forms:

$$\begin{aligned}\delta &= \frac{1}{(n-k)}(x_n + \ldots + x_{k+1}) - \frac{1}{k}(x_1 + \ldots + x_k) \\ &= \frac{n}{k}\{(x_n + \ldots + x_{k+1})/(n-k) - \bar{x}\} \\ &= \frac{n}{(n-k)}\{\bar{x} - (x_1 + \ldots + x_k)/k\}.\end{aligned} \tag{35}$$

When $k = (n-1)$ or 1, the δ of (35) reduces to the extreme deviate $x_n - \bar{x}$ or $\bar{x} - x_1$, multiplied by the factor $n/(n-1)$.

To obtain the distribution of δ of (34), we start off by introducing the following variates:

$$\left.\begin{aligned}\bar{x} &= \frac{1}{n}(x_1 + \ldots + x_n), \\ \gamma &= \frac{(x_n + \ldots + x_{n-l+1}) + (x_1 + \ldots + x_k)}{(k+l)} - \frac{(x_{k+1} + \ldots + x_{n-l})}{n-k-l}, \\ u_i &= \{i x_{i+1} - (x_i + x_{i-1} + \ldots + x_1)\} \quad (i = 1, \ldots, (k-1)), \\ v_i &= \{i x_{i+k+1} - (x_{i+k} + \ldots + x_{k+1})\} \quad (i = 1, \ldots, (n-k-l-1)), \\ w_i &= \{(x_n + \ldots + x_{n-i+1}) - i x_{n-i}\} \quad (i = 1, \ldots, (l-1)).\end{aligned}\right\} \tag{36}$$

Dividing the n new variates, $\delta, \bar{x}, \gamma, u_i, v_i, w_i$ by the square root of the sum of the coefficients of x on the right-hand side, we get the following orthogonal system of transformed variates

$$\left.\begin{aligned}
\delta' &= \sqrt{\frac{kl}{(k+l)}}\,\delta, \\
\bar{x}' &= \sqrt{n}\,\bar{x}, \\
\gamma' &= \sqrt{\frac{(k+l)(n-k-l)}{n}}\,\gamma, \\
u_i' &= \frac{u_i}{\sqrt{[i(i+1)]}} \quad (i = 1, \ldots, (k-1)), \\
v_i' &= \frac{v_i}{\sqrt{[i(i+1)]}} \quad (i = 1, \ldots, (n-k-l-1)), \\
w_i' &= \frac{w_i}{\sqrt{[i(i+1)]}} \quad (i = 1, \ldots, (l-1)).
\end{aligned}\right\} \tag{37}$$

It follows that

$$\sum_1^n (x_i^2) = \bar{x}'^2 + \delta'^2 + \gamma'^2 + \sum_1^{k-1} u_i'^2 + \sum_1^{n-k-l-1} v_i'^2 + \sum_1^{l-1} w_i'^2$$

$$= n\bar{x}^2 + \frac{kl}{(k+l)}\delta^2 + \frac{(k+l)(n-k-l)}{n}\gamma^2 + \sum_1^{k-1} \frac{u_i^2}{i(i+1)} + \sum_1^{n-k-l-1} \frac{v_i^2}{i(i+1)} + \sum_1^{l-1} \frac{w_i^2}{i(i+1)}, \tag{38}$$

$$\prod_1^n dx_i = d\bar{x}'\, d\delta'\, d\gamma' \prod_1^{k-1} du_i' \prod_1^{n-k-l-1} dv_i' \prod_1^{l-1} dw_i'$$

$$= \frac{d\bar{x}\, d\delta\, d\gamma \prod_1^{k-1} du_i \prod_1^{n-k-l-1} dv_i \prod_1^{l-1} dw_i}{(k-1)!\,(l-1)!\,(n-k-l-1)!} \tag{39}$$

The joint distribution of $\bar{x}, \delta, \gamma, u_i, v_i, w_i$ can be obtained from the joint distribution (10) of $x_1, \ldots, x_n$ in the form

$$\frac{n!(2\pi)^{-\frac{1}{2}n}}{(k-1)!\,(l-1)!\,(n-k-l-1)!}$$

$$\times \exp\left[-\frac{1}{2}\left\{n\bar{x}^2 + \frac{kl}{(k+l)}\delta^2 + \frac{(k+l)(n-k-l)}{n}\gamma^2 + \sum_1^{k-1} \frac{u_i^2}{i(i+1)} + \sum_1^{n-k-l-1} \frac{v_i^2}{i(i+1)} + \sum_1^{l-1} \frac{w_i^2}{i(i+1)}\right\}\right]$$

$$\times d\bar{x}\, d\delta\, d\gamma \prod_1^{k-1} du_i \prod_1^{n-k-l-1} dv_i \prod_1^{l-1} dw_i. \tag{40}$$

Integrating out for $\bar{x}$ between the limits $\pm\infty$, the joint distribution of $\delta, \gamma, u_i, v_i, w_i$ is

$$\frac{n!\,n^{-\frac{1}{2}}(2\pi)^{-\frac{1}{2}(n-1)}}{(k-1)!\,(l-1)!\,(n-k-l-1)!}$$

$$\times \exp\left[-\frac{1}{2}\left\{\frac{kl}{k+l}\delta^2 + \frac{(k+l)(n-k-l)}{n}\gamma^2 + \sum_1^{k-1} \frac{u_i^2}{i(i+1)} + \sum_1^{n-k-l-1} \frac{v_i^2}{i(i+1)} + \sum_1^{l-1} \frac{w_i^2}{i(i+1)}\right\}\right]$$

$$\times d\delta\, d\gamma \prod_1^{k-1} du_i \prod_1^{n-k-l-1} dv_i \prod_1^{l-1} dw_i. \tag{41}$$

To integrate out for γ, u_i, v_i and w_i we have to note that corresponding to the inequalities $x_1 \leqslant \ldots \leqslant x_n$ we now have

$$\left.\begin{aligned} &0 \leqslant u_1 \leqslant \ldots \leqslant u_{k-1}, \\ &0 \leqslant v_1 \leqslant \ldots \leqslant v_{n-k-l-1}, \\ &0 \leqslant w_1 \leqslant \ldots \leqslant w_{l-1}, \\ &\frac{1}{k}u_{k-1} + \sum_1^{n-k-l-1} \frac{v_i}{i(i+1)} \leqslant \frac{l\delta}{(k+l)} - \gamma, \\ &\frac{1}{l}w_{l-1} + \frac{1}{(n-k-l)}v_{n-k-l-1} \leqslant \frac{k\delta}{(k+l)} + \gamma. \end{aligned}\right\} \tag{42}$$

It is extremely difficult to carry out the integration for γ, u_i, v_i and w_i subject to the inequalities (42). In the following special cases, however, the last two inequalities in (42) take a simpler form and the integration can be carried forward to a certain stage in terms of G-functions.

Case I. $k+l=n$. Here γ and v_i disappear and the inequalities (42) reduce to

$$\left.\begin{aligned} &0 \leqslant u_1 \leqslant \ldots \leqslant u_{k-1}, \\ &0 \leqslant w_1 \leqslant \ldots \leqslant w_{l-1}, \\ &\frac{1}{k}u_{k-1} + \frac{1}{l}w_{l-1} \leqslant \delta. \end{aligned}\right\} \tag{43}$$

The joint distribution of δ, u_i and w_i is

$$\frac{n!\,n^{-\frac{1}{2}}(2\pi)^{-\frac{1}{2}(n-1)}}{(k-1)!\,(l-1)!} \exp\left[-\frac{1}{2}\left\{\frac{kl}{n}\delta^2 + \sum_1^{k-1}\frac{u_i^2}{i(i+1)} + \sum_1^{l-1}\frac{w_i^2}{i(i+1)}\right\}\right] d\delta \prod_1^{k-1} du_i \prod_1^{l-1} dw_i. \tag{44}$$

Putting $u_{k-1} = \epsilon$ and writing the last inequality in (43) as $w_{l-1} \leqslant l(\delta - \epsilon/k)$ we can integrate out u_i and w_i from (44) and obtain the distribution of δ in the form

$$f(\delta) = \frac{n!\,n^{-\frac{1}{2}}(2\pi)^{-\frac{1}{2}(n-1)}}{(k-1)!\,(l-1)!} \exp\left[-\frac{kl}{2n}\delta^2\right] \int_0^{k\delta} G_{l-1}\{l(\delta - \epsilon/k)\}\, G'_{k-1}(\epsilon)\, d\epsilon, \tag{45}$$

where $$G'_{k-1}(\epsilon) = \frac{d}{d\epsilon} G_{k-1}(\epsilon) = \exp\left[-\frac{\epsilon^2}{2k(k-1)}\right] G_{k-2}(\epsilon).$$

If we had written $w_{l-1} = \epsilon$ and $u_{k-1} \leqslant k(\delta - \epsilon/l)$, we should get an alternate form which is the same as (45) with k and l interchanged.

Putting $l = 1$, and $\delta = nu/(n-1)$ in (45) we get the distribution of the extreme deviate $x_n - \bar{x}$ given in (19).

Putting $l = 2$, which is appropriate for the case of two outlying observations in the same direction, the distribution of δ takes the form

$$f(\delta) = \frac{2(n-1)(n-2)\sqrt{n}}{(\sqrt{2\pi})^{n-1}} \exp\left[-\frac{(n-2)}{n}\delta^2\right] \int_0^{\delta} G_{n-3}\{(n-2)\theta\}\, e^{-(\delta-\theta)^2}\, d\theta. \tag{46}$$

If $k = l = m$ and $n = 2m$, δ becomes twice the mean deviation from the median, whose distribution has been obtained by Godwin (unpublished).

Case II. $k+l=n-1$. Here v_i's disappear and the inequalities (42) reduce to

$$\left.\begin{aligned} &0 \leqslant u_1 \leqslant \ldots \leqslant u_{k-1} \leqslant k\left(\frac{l\delta}{n-1} - \gamma\right), \\ &0 \leqslant w_1 \leqslant \ldots \leqslant w_{l-1} \leqslant l\left(\frac{k\delta}{n-1} + \gamma\right). \end{aligned}\right\} \tag{47}$$

The joint distribution of δ, γ, u_i's and w_i's is

$$\frac{n!\,n^{-\frac{1}{2}}(2\pi)^{-\frac{1}{2}(n-1)}}{(k-1)!\,(l-1)!}\exp\left[-\frac{1}{2}\left\{\frac{kl}{(n-1)}\delta^2+\frac{(n-1)}{n}\gamma^2+\sum_1^{k-1}\frac{u_i^2}{i(i+1)}+\sum_1^{l-1}\frac{w_i^2}{i(i+1)}\right\}\right]$$
$$\times d\delta\,d\gamma\prod_1^{k-1}du_i\prod_1^{l-1}dw_i. \tag{48}$$

Integrating out γ, u_i and w_i from (48) using the inequalities (47) we get the distribution of δ in the form

$$f(\delta)=\frac{n!\,n^{-\frac{1}{2}}(2\pi)^{-\frac{1}{2}(n-1)}}{(k-1)!\,(l-1)!}\exp\left[-\frac{kl}{2(n-1)}\delta^2\right]$$
$$\times\int_{-k\delta/(n-1)}^{l\delta/(n-1)}G_{k-1}\left\{k\left(\frac{l\delta}{n-1}-\gamma\right)\right\}G_{l-1}\left\{l\left(\frac{k\delta}{n-1}+\gamma\right)\right\}\exp\left[-\frac{(n-1)}{2n}\gamma^2\right]d\gamma. \tag{49}$$

Putting $k=l=m$ and $n=2m+1$, δ becomes n/m times the mean deviation from the median whose distribution has been obtained by Godwin (unpublished). If $m=1$, in particular, δ becomes the range in a sample of 3 and we get

$$f(\delta)=\frac{\sqrt{3}}{\pi}e^{-\frac{1}{2}\delta^2}\int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta}e^{-\gamma^2/3}d\gamma$$
$$=\frac{2\sqrt{3}}{\pi}e^{-\frac{1}{2}\delta^2}\int_0^{\frac{1}{2}\delta}e^{-\gamma^2/3}d\gamma, \tag{50}$$

which is the form given by McKay & Pearson (1933).

Case III. $k+l=n-2$. The inequalities become

$$\left.\begin{aligned}&0\leqslant u_1\leqslant\ldots\leqslant u_{k-1}\leqslant k\left(\frac{l\delta}{n-2}-\gamma-\tfrac{1}{2}v_1\right),\\&0\leqslant v_1,\\&0\leqslant w_1\leqslant\ldots\leqslant w_{l-1}\leqslant l\left(\frac{k\delta}{n-2}+\gamma-\tfrac{1}{2}v_1\right).\end{aligned}\right\} \tag{51}$$

The distribution of δ comes out in the form of a double integral involving product of two G-functions

$$f(\delta)=\frac{n!\,n^{-\frac{1}{2}}(2\pi)^{-\frac{1}{2}(n-1)}}{(k-1)!\,(l-1)!}\exp\left[-\frac{kl}{2(n-2)}\delta^2\right]$$
$$\times\int_0^{\delta}\int_{-\left(\frac{k\delta}{n-2}-\frac{1}{2}v_1\right)}^{\left(\frac{l\delta}{n-2}-\frac{1}{2}v_1\right)}G_{k-1}\left\{k\left(\frac{l\delta}{n-2}-\gamma-\tfrac{1}{2}v_1\right)\right\}G_{l-1}\left\{l\left(\frac{k\delta}{n-2}+\gamma-\tfrac{1}{2}v_1\right)\right\}\exp\left[-\left(\frac{v_1^2}{4}+\frac{(n-2)\gamma^2}{n}\right)\right]d\gamma\,dv_1. \tag{52}$$

If $n=2m$ and $k=l=m-1$, δ becomes a special case of Jones's statistic. In particular, if $m=2$, this reduces to the range in samples of 4 and its distribution becomes

$$\frac{6\sqrt{2}}{\pi\sqrt{\pi}}e^{-\frac{1}{2}\delta^2}\int_0^{\delta}e^{-\frac{1}{2}v_1^2}\int_0^{\frac{1}{2}(\delta-v_1)}e^{-\frac{1}{2}x^2}dx\,dv_1. \tag{53}$$

I should like to acknowledge warmly the help and guidance I have received from Prof. Pearson and Dr Hartley in the course of this investigation.

My thanks are also due to Dr E. T. Goodwin and Mr T. Vickers of the Mathematics Division of the National Physical Laboratory for their expert help in carrying out the calculations required for Tables 1 and 5.

Table 1. *Probability integral of the extreme deviate* $(x_n - \bar{x})/\sigma$ *or* $(\bar{x} - x_1)/\sigma$ *in 'normal' samples of size* n

u \ n	3	4	5	6	7	8	9
	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}
0·00							
0·01	124						
0·02	496	10					
0·03	1 115	36					
0·04	1 982	87	3				
0·05	3 095	169	9				
0·06	4 453	293	19	1			
0·07	6 056	464	35	2			
0·08	7 901	690	59	5	1		
0·09	9 987	979	95	9	1		
0·10	12 312	1 338	145	16	2		
0·11	14 874	1 777	212	25	3		
0·12	17 671	2 301	299	39	5	1	
0·13	20 701	2 917	410	57	8	1	
0·14	23 960	3 631	549	83	13	2	
0·15	27 446	4 451	720	116	19	3	
0·16	31 156	5 382	927	159	28	5	
0·17	35 086	6 430	1 175	214	39	7	1
0·18	39 234	7 601	1 468	283	55	10	2
0·19	43 595	8 901	1 812	368	75	15	3
0·20	48 166	10 334	2 210	472	101	21	5
0·21	52 943	11 906	2 669	597	134	29	7
0·22	57 922	13 621	3 192	747	175	41	10
0·23	63 099	15 483	3 786	924	225	55	13
0·24	68 469	17 495	4 455	1 132	288	73	18
0·25	74 028	19 663	5 204	1 375	363	96	25
0·26	79 772	21 988	6 039	1 656	453	124	34
0·27	85 696	24 474	6 964	1 978	561	159	45
0·28	91 794	27 124	7 985	2 346	689	202	59
0·29	98 063	29 939	9 106	2 765	838	254	77
0·30	104 497	32 922	10 333	3 237	1 013	317	99
0·31	111 091	36 075	11 670	3 768	1 215	392	126
0·32	117 840	39 398	13 121	4 361	1 448	480	159
0·33	124 738	42 893	14 691	5 021	1 715	585	200
0·34	131 781	46 560	16 384	5 754	2 018	707	248
0·35	138 963	50 400	18 205	6 562	2 363	850	306
0·36	146 279	54 412	20 157	7 451	2 752	1 015	375
0·37	153 723	58 597	22 244	8 426	3 188	1 205	456
0·38	161 290	62 954	24 469	9 490	3 677	1 423	551
0·39	168 975	67 481	26 835	10 649	4 221	1 672	662
0·40	176 771	72 179	29 346	11 906	4 825	1 954	791
0·41	184 674	77 045	32 004	13 266	5 493	2 273	940
0·42	192 677	82 079	34 813	14 734	6 228	2 631	1 111
0·43	200 777	87 277	37 773	16 313	7 036	3 033	1 307
0·44	208 966	92 638	40 887	18 007	7 921	3 482	1 530
0·45	217 239	98 159	44 157	19 821	8 886	3 981	1 783
0·46	225 591	103 838	47 584	21 758	9 936	4 535	2 069
0·47	234 017	109 672	51 169	23 822	11 076	5 147	2 390
0·48	242 512	115 658	54 913	26 015	12 309	5 820	2 751
0·49	251 069	121 793	58 816	28 342	13 640	6 560	3 154
0·50	259 684	128 073	62 880	30 805	15 073	7 370	3 602

Table 1 (cont.). *Probability integral of the extreme deviate* $(x_n - \bar{x})/\sigma$ *or* $(\bar{x} - x_1)/\sigma$ *in 'normal' samples of size* n

u \ n	3	4	5	6	7	8	9
	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}
0·50	259 684	128 073	62 880	30 805	15 073	7 370	3 602
0·51	268 351	134 495	67 103	33 407	16 611	8 254	4 099
0·52	277 066	141 054	71 487	36 151	18 259	9 216	4 649
0·53	285 823	147 748	76 029	39 039	20 021	10 260	5 255
0·54	294 616	154 572	80 730	42 073	21 899	11 391	5 922
0·55	303 442	161 522	85 589	45 255	23 899	12 612	6 652
0·56	312 295	168 594	90 604	48 587	26 023	13 928	7 450
0·57	321 170	175 782	95 774	52 069	28 275	15 343	8 321
0·58	330 063	183 084	101 097	55 705	30 657	16 860	9 267
0·59	338 969	190 494	106 572	59 493	33 173	18 484	10 294
0·60	347 883	198 008	112 195	63 436	35 825	20 218	11 404
0·61	356 801	205 621	117 965	67 534	38 616	22 066	12 602
0·62	365 720	213 328	123 879	71 786	41 549	24 032	13 893
0·63	374 633	221 124	129 935	76 193	44 625	26 118	15 279
0·64	383 538	229 006	136 129	80 754	47 846	28 329	16 766
0·65	392 430	236 967	142 459	85 468	51 214	30 668	18 356
0·66	401 305	245 003	148 921	90 334	54 730	33 137	20 054
0·67	410 159	253 110	155 511	95 352	58 396	35 740	21 863
0·68	418 989	261 283	162 227	100 521	62 213	38 478	23 787
0·69	427 791	269 516	169 064	105 838	66 180	41 355	25 830
0·70	436 562	277 805	176 019	111 302	70 300	44 373	27 995
0·71	445 298	286 145	183 088	116 912	74 572	47 534	30 285
0·72	453 996	294 532	190 266	122 666	78 996	50 839	32 703
0·73	462 652	302 960	197 550	128 560	83 571	54 290	35 253
0·74	471 264	311 425	204 935	134 593	88 298	57 889	37 936
0·75	479 829	319 923	212 417	140 761	93 176	61 637	40 756
0·76	488 344	328 448	219 992	147 062	98 203	65 535	43 715
0·77	496 805	336 997	227 655	153 493	103 380	69 583	46 814
0·78	505 211	345 564	235 402	160 050	108 703	73 781	50 057
0·79	513 559	354 147	243 228	166 731	114 172	78 131	53 444
0·80	521 847	362 739	251 129	173 531	119 785	82 632	56 978
0·81	530 072	371 337	259 100	180 447	125 540	87 284	60 660
0·82	538 232	379 938	267 137	187 476	131 434	92 086	64 490
0·83	546 325	388 536	275 235	194 613	137 466	97 038	68 471
0·84	554 349	397 129	283 390	201 855	143 632	102 138	72 601
0·85	562 303	405 711	291 597	209 197	149 930	107 386	76 883
0·86	570 184	414 280	299 851	216 636	156 357	112 780	81 315
0·87	577 991	422 831	308 149	224 167	162 910	118 319	85 899
0·88	585 722	431 362	316 484	231 786	169 585	124 000	90 633
0·89	593 376	439 868	324 854	239 488	176 380	129 822	95 517
0·90	600 951	448 346	333 254	247 270	183 290	135 783	100 550
0·91	608 446	456 793	341 679	255 127	190 313	141 880	105 732
0·92	615 860	465 206	350 125	263 054	197 444	148 111	111 061
0·93	623 192	473 581	358 588	271 047	204 680	154 472	116 536
0·94	630 441	481 917	367 063	279 102	212 016	160 961	122 154
0·95	637 606	490 210	375 547	287 214	219 449	167 575	127 913
0·96	644 686	498 457	384 035	295 379	226 974	174 310	133 811
0·97	651 680	506 656	392 523	303 591	234 587	181 164	139 848
0·98	658 588	514 803	401 008	311 848	242 285	188 132	146 021
0·99	665 408	522 895	409 486	320 143	250 062	195 211	152 329
1·00	672 142	530 930	417 952	328 474	257 914	202 397	158 771

Table 1 (cont.). *Probability integral of the extreme deviate* $(x_n - \bar{x})/\sigma$ *or* $(\bar{x} - x_1)/\sigma$ *in 'normal' samples of size n*

u \ n	3	4	5	6	7	8	9
	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}
1·00	672 142	530 930	417 952	328 474	257 914	202 397	158 771
1·01	678 787	538 909	426 404	336 835	265 838	209 687	165 335
1·02	685 344	546 827	434 838	345 222	273 827	217 076	172 023
1·03	691 812	554 683	443 250	353 632	281 879	224 561	178 832
1·04	698 191	562 474	451 636	362 059	289 989	232 136	185 758
1·05	704 481	570 199	459 994	370 499	298 151	239 799	192 798
1·06	710 682	577 856	468 321	378 949	306 362	247 544	199 947
1·07	716 793	585 443	476 613	387 405	314 618	255 368	207 203
1·08	722 815	592 959	484 867	395 861	322 913	263 266	214 561
1·09	728 748	600 402	493 080	404 316	331 244	271 234	222 018
1·10	734 592	607 770	501 250	412 764	339 606	279 267	229 568
1·11	740 346	615 064	509 374	421 202	347 995	287 360	237 209
1·12	746 011	622 280	517 449	429 626	356 406	295 510	244 935
1·13	751 588	629 418	525 472	438 032	364 835	303 712	252 743
1·14	757 077	636 478	533 442	446 418	373 278	311 961	260 627
1·15	762 477	643 458	541 356	454 780	381 732	320 253	268 585
1·16	767 790	650 356	549 212	463 114	390 191	328 583	276 610
1·17	773 015	657 173	557 007	471 417	398 652	336 948	284 700
1·18	778 154	663 908	564 740	479 686	407 110	345 342	292 848
1·19	783 206	670 559	572 408	487 918	415 563	353 761	301 052
1·20	788 172	677 127	580 010	496 110	424 006	362 202	309 305
1·21	793 053	683 611	587 545	504 259	432 435	370 659	317 604
1·22	797 849	690 010	595 009	512 363	440 847	379 128	325 945
1·23	802 561	696 324	602 403	520 419	449 239	387 606	334 322
1·24	807 190	702 552	609 724	528 423	457 606	396 088	342 731
1·25	811 735	708 695	616 971	536 375	465 946	404 570	351 169
1·26	816 199	714 752	624 143	544 271	474 255	413 049	359 629
1·27	820 581	720 723	631 238	552 109	482 530	421 520	368 109
1·28	824 882	726 608	638 256	559 887	490 767	429 979	376 604
1·29	829 104	732 407	645 195	567 603	498 965	438 424	385 109
1·30	833 246	738 120	652 055	575 255	507 120	446 849	393 621
1·31	837 309	743 746	658 834	582 841	515 229	455 252	402 135
1·32	841 295	749 287	665 531	590 359	523 290	463 629	410 647
1·33	845 205	754 743	672 147	597 808	531 300	471 978	419 153
1·34	849 038	760 112	678 680	605 185	539 256	480 293	427 650
1·35	852 796	765 397	685 131	612 491	547 156	488 573	436 133
1·36	856 479	770 596	691 497	619 722	554 998	496 814	444 599
1·37	860 090	775 711	697 779	626 878	562 780	505 014	453 044
1·38	863 628	780 742	703 977	633 958	570 499	513 168	461 465
1·39	867 094	785 688	710 089	640 960	578 154	521 276	469 858
1·40	870 489	790 552	716 117	647 884	585 742	529 333	478 219
1·41	873 814	795 333	722 059	654 727	593 262	537 337	486 546
1·42	877 071	800 031	727 916	661 490	600 712	545 286	494 835
1·43	880 259	804 647	733 688	668 172	608 090	553 178	503 084
1·44	883 381	809 183	739 374	674 771	615 396	561 010	511 288
1·45	886 436	813 637	744 975	681 288	622 627	568 779	519 446
1·46	889 425	818 012	750 490	687 722	629 782	576 485	527 554
1·47	892 351	822 307	755 920	694 071	636 859	584 124	535 610
1·48	895 213	826 524	761 265	700 336	643 859	591 695	543 611
1·49	898 012	830 663	766 525	706 516	650 779	599 197	551 556
1·50	900 750	834 724	771 701	712 611	657 619	606 627	559 440

Table 1 (cont.). *Probability integral of the extreme deviate* $(x_n - \bar{x})/\sigma$ *or* $(\bar{x} - x_1)/\sigma$ *in 'normal' samples by size n*

u \ n	3	4	5	6	7	8	9
	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}
1·50	900 750	834 724	771 701	712 611	657 619	606 627	559 440
1·51	903 428	838 710	776 792	718 621	664 377	613 983	567 262
1·52	906 045	842 619	781 800	724 545	671 053	621 265	575 021
1·53	908 604	846 453	786 724	730 383	677 647	628 471	582 713
1·54	911 105	850 214	791 565	736 136	684 156	635 600	590 337
1·55	913 550	853 901	796 324	741 803	690 582	642 650	597 892
1·56	915 938	857 515	801 000	747 384	696 923	649 620	605 374
1·57	918 271	861 058	805 595	752 879	703 179	656 509	612 783
1·58	920 550	864 531	810 109	758 289	709 349	663 317	620 117
1·59	922 775	867 933	814 543	763 613	715 433	670 042	627 375
1·60	924 949	871 266	818 897	768 852	721 431	676 684	634 555
1·61	927 071	874 531	823 172	774 007	727 343	683 241	641 656
1·62	929 142	877 728	827 368	779 076	733 169	689 714	648 677
1·63	931 164	880 859	831 487	784 062	738 908	696 101	655 616
1·64	933 137	883 924	835 529	788 964	744 560	702 402	662 473
1·65	935 062	886 925	839 494	793 783	750 126	708 617	669 246
1·66	936 940	889 862	843 384	798 518	755 605	714 745	675 935
1·67	938 772	892 737	847 199	803 172	760 999	720 786	682 540
1·68	940 559	895 549	850 941	807 743	766 306	726 740	689 058
1·69	942 301	898 300	854 609	812 233	771 527	732 607	695 491
1·70	944 000	900 991	858 205	816 643	776 663	738 387	701 837
1·71	945 656	903 623	861 729	820 973	781 714	744 078	708 095
1·72	947 270	906 197	865 182	825 223	786 680	749 683	714 266
1·73	948 843	908 713	868 566	829 395	791 561	755 200	720 349
1·74	950 376	911 172	871 881	833 488	796 359	760 630	726 344
1·75	951 870	913 577	875 128	837 505	801 073	765 973	732 251
1·76	953 324	915 926	878 308	841 445	805 704	771 229	738 069
1·77	954 741	918 222	881 421	845 309	810 253	776 398	743 799
1·78	956 121	920 465	884 469	849 099	814 720	781 482	749 440
1·79	957 464	922 656	887 453	852 814	819 106	786 479	754 993
1·80	958 772	924 795	890 373	856 456	823 411	791 391	760 458
1·81	960 045	926 885	893 230	860 025	827 637	796 219	765 834
1·82	961 284	928 925	896 026	863 523	831 783	800 961	771 123
1·83	962 489	930 917	898 760	866 950	835 851	805 620	776 324
1·84	963 662	932 861	901 435	870 308	839 842	810 196	781 438
1·85	964 803	934 759	904 051	873 596	843 756	814 689	786 466
1·86	965 913	936 610	906 608	876 816	847 593	819 099	791 407
1·87	966 992	938 417	909 109	879 969	851 356	823 429	796 261
1·88	968 042	940 180	911 553	883 056	855 044	827 677	801 031
1·89	969 062	941 899	913 942	886 077	858 658	831 846	805 715
1·90	970 054	943 576	916 276	889 034	862 200	835 935	810 315
1·91	971 018	945 211	918 557	891 927	865 670	839 945	814 832
1·92	971 954	946 805	920 785	894 757	869 068	843 878	819 265
1·93	972 864	948 359	922 962	897 526	872 397	847 734	823 616
1·94	973 748	949 874	925 087	900 234	875 657	851 514	827 885
1·95	974 607	951 350	927 163	902 882	878 848	855 219	832 073
1·96	975 441	952 789	929 189	905 471	881 972	858 849	836 181
1·97	976 251	954 190	931 168	908 002	885 029	862 405	840 209
1·98	977 037	955 556	933 099	910 476	888 021	865 889	844 159
1·99	977 800	956 886	934 983	912 894	890 949	869 301	848 031
2·00	978 541	958 181	936 822	915 257	893 813	872 642	851 826

Table 1 (cont.). *Probability integral of the extreme deviate* $(x_n - \bar{x})/\sigma$ *or* $(\bar{x} - x_1)/\sigma$ *in 'normal' samples of size* n

u \ n	3	4	5	6	7	8	9
	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}
2·00	978 541	958 181	936 822	915 257	893 813	872 642	851 826
2·01	979 260	959 442	938 616	917 566	896 614	875 913	855 544
2·02	979 958	960 670	940 366	919 821	899 353	879 116	859 188
2·03	980 635	961 866	942 073	922 023	902 032	882 250	862 757
2·04	981 291	963 030	943 738	924 174	904 650	885 317	866 252
2·05	981 928	964 162	945 362	926 274	907 210	888 318	869 675
2·06	982 545	965 265	946 945	928 325	909 712	891 253	873 027
2·07	983 144	966 337	948 488	930 326	912 157	894 124	876 307
2·08	983 724	967 380	949 992	932 280	914 545	896 932	879 518
2·09	984 286	968 395	951 458	934 187	916 879	899 678	882 661
2·10	984 832	969 382	952 886	936 047	919 158	902 362	885 735
2·11	985 360	970 342	954 278	937 862	921 384	904 986	888 742
2·12	985 872	971 275	955 633	939 632	923 558	907 550	891 684
2·13	986 367	972 182	956 954	941 359	925 680	910 056	894 561
2·14	986 847	973 064	958 240	943 043	927 752	912 504	897 373
2·15	987 312	973 921	959 493	944 685	929 775	914 896	900 123
2·16	987 763	974 754	960 712	946 286	931 748	917 232	902 811
2·17	988 198	975 563	961 899	947 846	933 674	919 514	905 438
2·18	988 620	976 349	963 055	949 367	935 553	921 741	908 005
2·19	989 029	977 113	964 180	950 850	937 386	923 916	910 513
2·20	989 424	977 855	965 274	952 294	939 173	926 040	912 963
2·21	989 806	978 575	966 339	953 701	940 917	928 112	915 355
2·22	990 176	979 275	967 375	955 072	942 617	930 134	917 692
2·23	990 534	979 954	968 383	956 407	944 274	932 107	919 974
2·24	990 880	980 613	969 364	957 708	945 890	934 032	922 201
2·25	991 214	981 253	970 317	958 974	947 465	935 910	924 375
2·26	991 538	981 874	971 244	960 207	949 000	937 742	926 498
2·27	991 851	982 476	972 145	961 407	950 496	939 528	928 568
2·28	992 153	983 061	973 022	962 575	951 953	941 270	930 589
2·29	992 445	983 628	973 873	963 712	953 373	942 968	932 560
2·30	992 727	984 178	974 701	964 819	954 756	944 623	934 483
2·31	992 999	984 711	975 505	965 896	956 102	946 236	936 358
2·32	993 263	985 228	976 287	966 943	957 414	947 808	938 187
2·33	993 517	985 730	977 046	967 962	958 691	949 340	939 970
2·34	993 763	986 216	977 784	968 953	959 934	950 832	941 708
2·35	994 000	986 687	978 500	969 916	961 144	952 286	943 402
2·36	994 229	987 144	979 196	970 853	962 322	953 702	945 053
2·37	994 450	987 587	979 871	971 765	963 468	955 081	946 662
2·38	994 663	988 015	980 527	972 650	964 583	956 424	948 230
2·39	994 869	988 431	981 164	973 511	965 668	957 732	949 758
2·40	995 067	988 833	981 782	974 348	966 723	959 005	951 245
2·41	995 259	989 223	982 381	975 161	967 750	960 243	952 694
2·42	995 443	989 600	982 963	975 951	968 748	961 449	954 105
2·43	995 621	989 966	983 528	976 718	969 719	962 622	955 479
2·44	995 793	990 320	984 075	977 464	970 663	963 764	956 817
2·45	995 959	990 662	984 607	978 188	971 581	964 875	958 119
2·46	996 118	990 993	985 122	978 891	972 473	965 955	959 386
2·47	996 272	991 314	985 622	979 574	973 340	967 006	960 620
2·48	996 420	991 625	986 106	980 237	974 183	968 028	961 820
2·49	996 563	991 925	986 576	980 881	975 001	969 021	962 987
2·50	996 701	992 215	987 031	981 505	975 797	969 987	964 123

Table 1 (cont.). *Probability integral of the extreme deviate* $(x_n - \bar{x})/\sigma$ *or* $(\bar{x} - x_1)/\sigma$ *in 'normal' samples of size n*

u \ n	3	4	5	6	7	8	9
	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}
2·50	996 701	992 215	987 031	981 505	975 797	969 987	964 123
2·51	996 833	992 496	987 473	982 112	976 569	970 926	965 228
2·52	996 961	992 768	987 901	982 700	977 320	971 839	966 303
2·53	997 084	993 031	988 315	983 271	978 049	972 726	967 347
2·54	997 202	993 285	988 717	983 825	978 757	973 588	968 363
2·55	997 316	993 530	989 106	984 362	979 444	974 426	969 351
2·56	997 425	993 768	989 483	984 883	980 111	975 240	970 311
2·57	997 531	993 997	989 848	985 389	980 759	976 030	971 244
2·58	997 632	994 219	990 201	985 879	981 387	976 798	972 151
2·59	997 730	994 433	990 544	986 354	981 998	977 544	973 032
2·60	997 824	994 640	990 875	986 815	982 590	978 268	973 889
2·61	997 914	994 840	991 196	987 261	983 164	978 971	974 720
2·62	998 001	995 033	991 506	987 694	983 721	979 653	975 528
2·63	998 084	995 219	991 807	988 113	984 262	980 316	976 313
2·64	998 165	995 399	992 098	988 520	984 786	980 959	977 074
2·65	998 242	995 573	992 379	988 914	985 294	981 583	977 814
2·66	998 316	995 740	992 651	989 295	985 787	982 188	978 533
2·67	998 387	995 902	992 915	989 665	986 265	982 776	979 230
2·68	998 456	996 058	993 169	990 022	986 728	983 346	979 907
2·69	998 521	996 209	993 416	990 369	987 178	983 898	980 563
2·70	998 585	996 355	993 654	990 705	987 613	984 434	981 201
2·71	998 645	996 495	993 884	991 029	988 035	984 954	981 819
2·72	998 703	996 630	994 107	991 344	988 443	985 458	982 419
2·73	998 759	996 761	994 322	991 648	988 839	985 947	983 001
2·74	998 813	996 886	994 530	991 943	989 223	986 421	983 566
2·75	998 865	997 008	994 731	992 228	989 594	986 880	984 113
2·76	998 914	997 125	994 925	992 504	989 954	987 325	984 644
2·77	998 961	997 237	995 112	992 771	990 303	987 756	985 159
2·78	999 007	997 346	995 293	993 029	990 640	988 174	985 658
2·79	999 050	997 451	995 468	993 279	990 967	988 579	986 141
2·80	999 092	997 551	995 637	993 520	991 283	988 971	986 610
2·81	999 132	997 649	995 800	993 754	991 589	989 351	987 064
2·82	999 171	997 742	995 958	993 979	991 885	989 719	987 505
2·83	999 208	997 832	996 110	994 198	992 172	990 075	987 931
2 84	999 243	997 919	996 257	994 408	992 449	990 420	988 344
2·85	999 277	998 003	996 399	994 612	992 717	990 754	988 745
2·86	999 309	998 083	996 535	994 809	992 977	991 077	989 132
2·87	999 340	998 161	996 667	995 000	993 228	991 390	989 508
2·88	999 370	998 235	996 795	995 184	993 471	991 693	989 871
2·89	999 399	998 307	996 917	995 361	993 705	991 986	990 223
2·90	999 426	998 376	997 036	995 533	993 932	992 269	990 564
2·91	999 452	998 442	997 150	995 699	994 152	992 543	990 894
2·92	999 477	998 506	997 260	995 859	994 364	992 809	991 213
2·93	999 501	998 568	997 366	996 014	994 569	993 065	991 522
2·94	999 524	998 627	997 469	996 163	994 767	993 314	991 821
2·95	999 546	998 683	997 567	996 307	994 959	993 554	992 111
2·96	999 567	998 738	997 663	996 446	995 143	993 786	992 390
2·97	999 587	998 790	997 754	996 580	995 322	994 010	992 661
2·98	999 606	998 841	997 842	996 710	995 495	994 227	992 923
2·99	999 625	998 889	997 928	996 835	995 662	994 437	993 176
3·00	999 642	998 936	998 010	996 955	995 823	994 639	993 421

Table 1 (cont.). *Probability integral of the extreme deviate* $(x_n - \bar{x})/\sigma$ *or* $(\bar{x} - x_1)/\sigma$ *in 'normal' samples of size n*

u \ n	3	4	5	6	7	8	9
	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}
3·00	999 642	998 936	998 010	996 955	995 823	994 639	993 421
3·01	999 659	998 981	998 088	997 072	995 978	994 835	993 658
3·02	999 675	999 024	998 164	997 184	996 128	995 024	993 887
3·03	999 690	999 065	998 238	997 292	996 273	995 207	994 108
3·04	999 705	999 105	998 308	997 396	996 413	995 383	994 322
3·05	999 719	999 143	998 376	997 497	996 548	995 554	994 528
3·06	999 732	999 179	998 441	997 594	996 678	995 718	994 728
3·07	999 745	999 215	998 504	997 687	996 804	995 877	994 921
3·08	999 757	999 248	998 565	997 777	996 925	996 031	995 107
3·09	999 769	999 281	998 623	997 864	997 043	996 179	995 287
3·10	999 780	999 312	998 679	997 948	997 155	996 322	995 461
3·11	999 791	999 341	998 733	998 029	997 264	996 460	995 629
3·12	999 801	999 370	998 785	998 106	997 369	996 594	995 792
3·13	999 810	999 397	998 834	998 181	997 471	996 722	995 948
3·14	999 820	999 424	998 882	998 253	997 568	996 846	996 099
3·15	999 829	999 449	998 929	998 322	997 662	996 966	996 245
3·16	999 837	999 473	998 973	998 389	997 753	997 082	996 386
3·17	999 845	999 496	999 015	998 454	997 840	997 193	996 522
3·18	999 853	999 519	999 056	998 515	997 925	997 301	996 653
3·19	999 860	999 540	999 096	998 575	998 006	997 404	996 780
3·20	999 867	999 560	999 134	998 632	998 084	997 504	996 902
3·21	999 873	999 580	999 170	998 688	998 159	997 600	997 020
3·22	999 880	999 599	999 205	998 741	998 232	997 693	997 133
3·23	999 886	999 617	999 238	998 792	998 302	997 783	997 243
3·24	999 891	999 634	999 270	998 841	998 369	997 869	997 349
3·25	999 897	999 650	999 301	998 888	998 434	997 952	997 451
3·26	999 902	999 666	999 331	998 934	998 497	998 032	997 549
3·27	999 907	999 681	999 359	998 978	998 557	998 109	997 644
3·28	999 912	999 696	999 387	999 020	998 615	998 184	997 735
3·29	999 916	999 709	999 413	999 060	998 670	998 255	997 823
3·30	999 920	999 723	999 438	999 099	998 724	998 324	997 908
3·31	999 924	999 735	999 462	999 136	998 775	998 391	997 989
3·32	999 928	999 747	999 486	999 172	998 825	998 455	998 068
3·33	999 932	999 759	999 508	999 207	998 873	998 516	998 144
3·34	999 935	999 770	999 529	999 240	998 919	998 575	998 217
3·35	999 939	999 781	999 550	999 272	998 963	998 633	998 287
3·36	999 942	999 791	999 569	999 302	999 005	998 687	998 355
3·37	999 945	999 801	999 588	999 332	999 046	998 740	998 420
3·38	999 948	999 810	999 606	999 360	999 085	998 791	998 483
3·39	999 951	999 819	999 624	999 387	999 123	998 840	998 544
3·40	999 953	999 827	999 640	999 413	999 159	998 887	998 602
3·41	999 956	999 835	999 656	999 438	999 194	998 932	998 658
3·42	999 958	999 843	999 671	999 462	999 228	998 976	998 712
3·43	999 960	999 850	999 686	999 485	999 260	999 018	998 764
3·44	999 962	999 857	999 700	999 507	999 291	999 058	998 813
3·45	999 964	999 864	999 713	999 528	999 320	999 097	998 862
3·46	999 966	999 871	999 726	999 549	999 349	999 134	998 908
3·47	999 968	999 877	999 738	999 568	999 376	999 169	998 952
3·48	999 970	999 883	999 750	999 587	999 403	999 204	998 995
3·49	999 971	999 888	999 761	999 605	999 428	999 237	999 036
3·50	999 973	999 894	999 772	999 622	999 452	999 269	999 076

Table 1 (cont.). *Probability integral of the extreme deviate* $(x_n - \bar{x})/\sigma$ *or* $(\bar{x} - x_1)/\sigma$ *in 'normal' samples of size n*

u \ n	3	4	5	6	7	8	9
	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}
3·50	999 973	999 894	999 772	999 622	999 452	999 269	999 076
3·51	999 974	999 899	999 783	999 638	999 476	999 299	999 114
3·52	999 976	999 904	999 792	999 654	999 498	999 328	999 150
3·53	999 977	999 908	999 802	999 669	999 519	999 357	999 135
3·54	999 978	999 913	999 811	999 684	999 540	999 384	999 219
3·55	999 979	999 917	999 820	999 698	999 560	999 410	999 251
3·56	999 981	999 921	999 828	999 711	999 579	999 435	999 282
3·57	999 982	999 925	999 836	999 724	999 597	999 458	999 312
3·58	999 983	999 929	999 843	999 736	999 614	999 481	999 341
3·59	999 984	999 932	999 851	999 748	999 631	999 503	999 369
3·60	999 985	999 935	999 858	999 759	999 647	999 525	999 395
3·61	999 985	999 939	999 864	999 770	999 662	999 545	999 421
3·62	999 986	999 942	999 870	999 780	999 677	999 564	999 445
3·63	999 987	999 945	999 876	999 790	999 691	999 583	999 469
3·64	999 988	999 947	999 882	999 800	999 705	999 601	999 491
3·65	999 988	999 950	999 888	999 809	999 718	999 618	999 513
3·66	999 989	999 952	999 893	999 817	999 730	999 635	999 533
3·67	999 990	999 955	999 898	999 826	999 742	999 651	999 553
3·68	999 990	999 957	999 903	999 834	999 754	999 666	999 573
3·69	999 991	999 959	999 908	999 841	999 765	999 680	999 591
3·70	999 991	999 961	999 912	999 848	999 775	999 694	999 609
3·71	999 992	999 963	999 916	999 855	999 785	999 708	999 625
3·72	999 992	999 965	999 920	999 862	999 795	999 720	999 642
3·73	999 993	999 967	999 924	999 868	999 804	999 733	999 657
3·74	999 993	999 968	999 928	999 874	999 813	999 745	999 672
3·75	999 994	999 970	999 931	999 880	999 821	999 756	999 686
3·76	999 994	999 972	999 934	999 886	999 829	999 767	999 700
3·77	999 994	999 973	999 938	999 891	999 837	999 777	999 713
3·78	999 995	999 974	999 941	999 896	999 844	999 787	999 726
3·79	999 995	999 976	999 943	999 901	999 852	999 796	999 738
3·80	999 995	999 977	999 946	999 906	999 858	999 806	999 749
3·81	999 995	999 978	999 949	999 910	999 865	999 814	999 760
3·82	999 996	999 979	999 951	999 914	999 871	999 823	999 771
3·83	999 996	999 980	999 954	999 918	999 877	999 831	999 781
3·84	999 996	999 981	999 956	999 922	999 883	999 838	999 791
3·85	999 996	999 982	999 958	999 926	999 888	999 846	999 800
3·86	999 997	999 983	999 960	999 929	999 893	999 852	999 809
3·87	999 997	999 984	999 962	999 933	999 898	999 859	999 817
3·88	999 997	999 985	999 964	999 936	999 903	999 866	999 826
3·89	999 997	999 986	999 966	999 939	999 907	999 872	999 834
3·90	999 997	999 987	999 968	999 942	999 912	999 878	999 841
3·91	999 997	999 987	999 969	999 945	999 916	999 883	999 848
3·92	999 998	999 988	999 971	999 947	999 920	999 889	999 855
3·93	999 998	999 989	999 972	999 950	999 924	999 894	999 862
3·94	999 998	999 989	999 974	999 952	999 927	999 899	999 868
3·95	999 998	999 990	999 975	999 955	999 931	999 903	999 874
3·96	999 998	999 990	999 976	999 957	999 934	999 908	999 880
3·97	999 998	999 991	999 977	999 959	999 937	999 912	999 885
3·98	999 998	999 991	999 979	999 961	999 940	999 916	999 890
3·99	999 998	999 992	999 980	999 963	999 943	999 920	999 895
4·00	999 999	999 992	999 981	999 965	999 946	999 924	999 900

Table 1 (cont.). *Probability integral of the extreme deviate* $(x_n - \bar{x})/\sigma$ *or* $(\bar{x} - x_1)/\sigma$ *in 'normal' samples of size* n

u \ n	3	4	5	6	7	8	9
	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}
4·00	999 999	999 992	999 981	999 965	999 946	999 924	999 900
4·01	999 999	999 993	999 982	999 966	999 948	999 927	999 905
4·02	999 999	999 993	999 983	999 968	999 951	999 931	999 909
4·03	999 999	999 993	999 983	999 970	999 953	999 934	999 913
4·04	999 999	999 994	999 984	999 971	999 955	999 937	999 917
4·05	999 999	999 994	999 985	999 973	999 958	999 940	999 921
4·06	999 999	999 994	999 986	999 974	999 960	999 943	999 925
4·07	999 999	999 995	999 987	999 975	999 962	999 946	999 928
4·08	999 999	999 995	999 987	999 976	999 964	999 948	999 932
4·09	999 999	999 995	999 988	999 978	999 965	999 951	999 935
4·10	999 999	999 996	999 989	999 979	999 967	999 953	999 938
4·11	999 999	999 996	999 989	999 980	999 969	999 955	999 941
4·12	999 999	999 996	999 990	999 981	999 970	999 957	999 944
4·13	999 999	999 996	999 990	999 982	999 972	999 959	999 946
4·14	999 999	999 996	999 991	999 983	999 973	999 961	999 949
4·15	1·000 000	999 997	999 991	999 984	999 975	999 963	999 951
4·16		999 997	999 992	999 984	999 976	999 965	999 954
4·17		999 997	999 992	999 985	999 977	999 967	999 956
4·18		999 997	999 993	999 986	999 978	999 968	999 958
4·19		999 997	999 993	999 987	999 979	999 970	999 960
4·20		999 997	999 993	999 987	999 980	999 971	999 962
4·21		999 998	999 994	999 988	999 981	999 973	999 964
4·22		999 998	999 994	999 989	999 982	999 974	999 965
4·23		999 998	999 994	999 989	999 983	999 975	999 967
4·24		999 998	999 995	999 990	999 984	999 977	999 969
4·25		999 998	999 995	999 990	999 985	999 978	999 970
4·26		999 998	999 995	999 991	999 986	999 979	999 971
4·27		999 998	999 996	999 991	999 986	999 980	999 973
4·28		999 998	999 996	999 992	999 987	999 981	999 974
4·29		999 998	999 996	999 992	999 988	999 982	999 975
4·30		999 999	999 996	999 993	999 988	999 983	999 977
4·31		999 999	999 996	999 993	999 989	999 983	999 978
4·32		999 999	999 997	999 993	999 990	999 984	999 979
4·33		999 999	999 997	999 994	999 990	999 985	999 980
4·34		999 999	999 997	999 994	999 991	999 986	999 981
4·35		999 999	999 997	999 994	999 991	999 987	999 982
4·36		999 999	999 997	999 995	999 992	999 987	999 983
4·37		999 999	999 997	999 995	999 992	999 988	999 983
4·38		999 999	999 998	999 995	999 992	999 988	999 984
4·39		999 999	999 998	999 995	999 993	999 989	999 985
4·40		999 999	999 998	999 996	999 993	999 990	999 986
4·41		999 999	999 998	999 996	999 994	999 990	999 986
4·42		999 999	999 998	999 996	999 994	999 991	999 987
4·43		999 999	999 998	999 996	999 994	999 991	999 988
4·44		999 999	999 998	999 996	999 995	999 992	999 988
4·45		999 999	999 998	999 997	999 995	999 992	999 989
4·46		999 999	999 998	999 997	999 995	999 992	999 989
4·47		999 999	999 999	999 997	999 996	999 993	999 990
4·48		999 999	999 999	999 997	999 996	999 993	999 991
4·49		999 999	999 999	999 997	999 996	999 993	999 991
4·50		999 999	999 999	999 998	999 996	999 994	999 992

Table 1 (cont.). *Probability integral of the extreme deviate* $(x_n - \bar{x})/\sigma$ *or* $(\bar{x} - x_1)/\sigma$ *in 'normal' samples of size n*

u \ n	3	4	5	6	7	8	9
	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}
4·50		999 999	999 999	999 998	999 996	999 994	999 992
4·51		999 999	999 999	999 998	999 997	999 994	999 992
4·52		999 999	999 999	999 998	999 997	999 994	999 992
4·53		999 999	999 999	999 998	999 997	999 995	999 993
4·54		999 999	999 999	999 998	999 997	999 995	999 993
4·55		1·000 000	999 999	999 998	999 997	999 995	999 993
4·56			999 999	999 998	999 997	999 996	999 994
4·57			999 999	999 998	999 998	999 996	999 994
4·58			999 999	999 998	999 998	999 996	999 994
4·59			999 999	999 998	999 998	999 996	999 995
4·60			1·000 000	999 999	999 998	999 996	999 995
4·61				999 999	999 998	999 996	999 995
4·62				999 999	999 998	999 997	999 995
4·63				999 999	999 998	999 997	999 995
4·64				999 999	999 998	999 997	999 996
4·65				999 999	999 999	999 997	999 996
4·66				999 999	999 999	999 997	999 996
4·67				999 999	999 999	999 997	999 996
4·68				999 999	999 999	999 998	999 997
4·69				999 999	999 999	999 998	999 997
4·70				999 999	999 999	999 998	999 997

Table 2. *Percentage points of the extreme deviate* $(x_n - \bar{x})/\sigma$ *or* $(\bar{x} - x_1)/\sigma$

Size of sample n	Lower percentage points						Upper percentage points					
	0·1	0·5	1·0	2·5	5·0	10·0	10·0	5·0	2·5	1·0	0·5	0·1
3	0·03	0·06	0·09	0·14	0·20	0·29	1·50	1·74	1·95	2·22	2·40	2·78
4	0·09	0·16	0·20	0·27	0·35	0·45	1·70	1·94	2·16	2·43	2·62	3·01
5	0·16	0·25	0·30	0·38	0·47	0·58	1·83	2·08	2·30	2·57	2·76	3·17
6	0·23	0·33	0·38	0·48	0·56	0·68	1·94	2·18	2·41	2·68	2·87	3·28
7	0·30	0·40	0·46	0·56	0·65	0·76	2·02	2·27	2·49	2·76	2·95	3·36
8	0·36	0·47	0·53	0·62	0·72	0·84	2·09	2·33	2·56	2·83	3·02	3·43
9	0·41	0·53	0·59	0·69	0·78	0·90	2·15	2·39	2·61	2·88	3·07	3·48

Table 5. *Auxiliary quantities required for calculating the probability integral of the studentized extreme deviate* $(x_n - \bar{x})/s_\nu$ *or* $(\bar{x} - x_1)/s_\nu$

n / Q	3			4			5			6			n / Q
	a_0	a_1	a_2	a_0	a_1	a_2	a_0	a_1	a_2	a_0	a_1	a_2	
0·00	0·000000	—	—	0·000000	—	—	0·000000	—	—	0·000000	—	—	0·00
0·20	0·048166	−0·00277	—	0·010334	+0·0063	−0·003	0·002210	+0·0039	−0·001	0·000472	+0·0016	—	0·20
0·40	0·176771	−0·03594	+0·012	0·072179	+0·0196	−0·022	0·029346	+0·0331	−0·028	0·011906	+0·0281	−0·016	0·40
0·60	0·347883	−0·13016	+0·076	0·198008	−0·0243	+0·004	0·112195	+0·0466	−0·076	0·063436	+0·0787	−0·113	0·60
0·80	0·521847	−0·26547	+0·195	0·362739	−0·1622	+0·169	0·251129	−0·0463	+0·038	0·173531	+0·0486	−0·120	0·80
1·00	0·672142	−0·38720	+0·275	0·530930	−0·3459	+0·402	0·417952	−0·2483	+0·381	0·328474	−0·1323	+0·231	1·00
1·20	0·788172	−0·45487	+0·235	0·677127	−0·4989	+0·524	0·580010	−0·4716	+0·721	0·496110	−0·3993	+0·783	1·20
1·40	0·870489	−0·45998	+0·083	0·790552	−0·5748	+0·430	0·716117	−0·6273	+0·792	0·647884	−0·6314	+1·103	1·40
1·60	0·924949	−0·41502	−0·131	0·871266	−0·5686	+0·148	0·818897	−0·6788	+0·530	0·768852	−0·7510	+0·953	1·60
1·80	0·958772	−0·34014	−0·359	0·924795	−0·5018	−0·215	0·890373	−0·6376	+0·059	0·856456	−0·7492	+0·426	1·80
2·00	0·978541	−0·25540	−0·551	0·958181	−0·4036	−0·549	0·936822	−0·5384	−0·441	0·915257	−0·6599	−0·237	2·00
2·20	0·989424	−0·17657	−0·659	0·977855	−0·2994	−0·787	0·965274	−0·4167	−0·835	0·952294	−0·5278	−0·810	2·20
2·40	0·995067	−0·11277	−0·666	0·988833	−0·2062	−0·891	0·981782	−0·2990	−1·058	0·974348	−0·3897	−1·176	2·40
2·60	0·997824	−0·06667	−0·587	0·994640	−0·1323	−0·869	0·990875	−0·2002	−1·107	0·986815	−0·2682	−1·314	2·60
2·80	0·999092	−0·03656	−0·460	0·997551	−0·0794	−0·752	0·995637	−0·1255	−1·014	0·993520	−0·1729	−1·255	2·80
3·00	0·999642	−0·01865	−0·323	0·998936	−0·0445	−0·587	0·998010	−0·0739	−0·836	0·996955	−0·1048	−1·072	3·00
3·20	0·999867	−0·00884	−0·207	0·999560	−0·0234	−0·417	0·999134	−0·0409	−0·625	0·998632	−0·0598	−0·833	3·20
3·40	0·999953	−0·00391	−0·119	0·999827	−0·0115	−0·271	0·999640	−0·0213	−0·432	0·999413	−0·0322	−0·594	3·40
3·60	0·999985	−0·00163	−0·063	0·999935	−0·0054	−0·161	0·999858	−0·0105	−0·276	0·999759	−0·0164	−0·392	3·60
3·80	0·999995	−0·00064	−0·031	0·999977	−0·0024	−0·091	0·999946	−0·0049	−0·162	0·999906	−0·0079	−0·243	3·80
4·00	0·999999	−0·00021	−0·014	0·999992	−0·0010	−0·045	0·999981	−0·0021	−0·090	0·999965	−0·0036	−0·136	4·00

Table 5 (cont.). *Auxiliary quantities required for calculating the probability integral of the studentized extreme deviate* $(x_n - \bar{x})/s_y$ *or* $(\bar{x} - x_1)/s_y$

n / Q	7			8			9			n / Q
	a_0	a_1	a_2	a_0	a_1	a_2	a_0	a_1	a_2	
0·00	0·000000	—	—	0·000000	—	—	0·000000	—	—	0·00
0·20	0·000101	+0·0005	+0·001	0·000021	+0·0002	—	0·000005	—	—	0·20
0·40	0·004825	+0·0192	−0·003	0·001954	+0·0117	+0·005	0·000791	+0·0066	+0·008	0·40
0·60	0·035825	+0·0842	−0·105	0·020218	+0·0758	−0·073	0·011404	+0·0619	−0·033	0·60
0·80	0·119785	+0·1133	−0·246	0·082632	+0·1498	−0·314	0·056978	+0·1642	−0·324	0·80
1·00	0·257914	−0·0199	+0·007	0·202397	+0·0775	−0·235	0·158771	+0·1552	−0·457	1·00
1·20	0·424006	−0·3018	+0·706	0·362202	−0·1934	+0·516	0·309305	−0·0838	+0·24	1·20
1·40	0·585742	−0·5988	+1·320	0·529333	−0·5395	+1·425	0·478219	−0·4615	+1·41	1·40
1·60	0·721431	−0·7907	+1·374	0·676684	−0·8028	+1·757	0·634555	−0·7919	+2·08	1·60
1·80	0·823411	−0·8384	+0·852	0·791391	−0·9071	+1·313	0·760458	−0·9573	+1·79	1·80
2·00	0·893813	−0·7685	+0·051	0·872642	−0·8648	+0·402	0·851826	−0·9493	+0·81	2·00
2·20	0·939173	−0·6327	−0·714	0·926040	−0·7312	−0·556	0·912963	−0·8234	−0·34	2·20
2·40	0·966723	−0·4780	−1·238	0·959005	−0·5635	−1·266	0·951245	−0·6462	−1·24	2·40
2·60	0·982590	−0·3356	−1·488	0·978268	−0·4022	−1·638	0·973889	−0·4678	−1·76	2·60
2·80	0·991283	−0·2206	−1·478	0·988971	−0·2684	−1·685	0·986610	−0·3160	−1·88	2·80
3·00	0·995823	−0·1364	−1·299	0·994639	−0·1684	−1·518	0·993421	−0·2006	−1·73	3·00
3·20	0·998084	−0·0795	−1·035	0·997504	−0·0997	−1·233	0·996902	−0·1201	−1·43	3·20
3·40	0·999159	−0·0437	−0·756	0·998887	−0·0558	−0·918	0·998602	−0·0679	−1·08	3·40
3·60	0·999647	−0·0229	−0·511	0·999525	−0·0295	−0·633	0·999395	−0·0365	−0·75	3·60
3·80	0·999858	−0·0112	−0·324	0·999806	−0·0147	−0·404	0·999749	−0·0185	−0·49	3·80
4·00	0·999946	−0·0052	−0·186	0·999924	−0·0070	−0·243	0·999900	−0·0090	−0·30	4·00
4·20	0·999980	−0·0024	−0·106	0·999971	−0·0032	−0·134	0·999962	−0·0042	−0·16	4·20
4·40	0·999993	−0·0010	−0·054	0·999990	−0·0014	−0·070	0·999986	−0·0018	−0·08	4·40
4·60	0·999998	−0·0004	−0·021	0·999996	−0·0006	−0·035	0·999995	−0·0007	−0·04	4·60

Table 6A. *Lower per cent points of the studentized extreme deviate* $(x_n - \bar{x})/s_\nu$ *or* $(\bar{x} - x_1)/s_\nu$

n / ν	5%							1%						
	3	4	5	6	7	8	9	3	4	5	6	7	8	9
10	0·20	0·35	0·46	0·55	0·62	0·69	0·74	0·09	0·19	0·29	0·37	0·43	0·49	0·54
15	0·20	0·35	0·46	0·55	0·63	0·70	0·75	0·09	0·19	0·29	0·37	0·44	0·50	0·56
30	0·20	0·35	0·46	0·56	0·64	0·70	0·77	0·09	0·20	0·29	0·38	0·45	0·51	0·57
∞	0·20	0·35	0·47	0·56	0·65	0·72	0·78	0·09	0·20	0·30	0·38	0·46	0·53	0·59

B. *Upper per cent points of the studentized extreme deviate* $(x_n - \bar{x})/s_\nu$ *or* $(\bar{x} - x_1)/s_\nu$

n / ν	5%							1%						
	3	4	5	6	7	8	9	3	4	5	6	7	8	9
10	2·02	2·29	2·49	2·63	2·75	2·85	2·93	2·76	3·05	3·25	3·39	3·50	3·59	3·67
11	1·99	2·26	2·44	2·58	2·70	2·79	2·87	2·71	3·00	3·19	3·33	3·44	3·53	3·61
12	1·97	2·22	2·40	2·54	2·65	2·75	2·83	2·67	2·95	3·14	3·28	3·39	3·48	3·55
13	1·95	2·20	2·38	2·51	2·62	2·71	2·79	2·63	2·91	3·10	3·24	3·34	3·43	3·51
14	1·93	2·18	2·35	2·48	2·59	2·68	2·76	2·60	2·87	3·06	3·20	3·30	3·39	3·47
15	1·92	2·16	2·33	2·46	2·56	2·65	2·73	2·57	2·84	3·02	3·16	3·27	3·35	3·43
16	1·90	2·14	2·31	2·44	2·54	2·63	2·70	2·55	2·81	3·00	3·13	3·24	3·32	3·39
17	1·89	2·13	2·30	2·42	2·52	2·61	2·68	2·52	2·79	2·97	3·10	3·21	3·29	3·36
18	1·88	2·12	2·28	2·41	2·51	2·59	2·66	2·50	2·77	2·95	3·08	3·18	3·27	3·34
19	1·87	2·11	2·27	2·39	2·49	2·58	2·65	2·49	2·75	2·92	3·06	3·16	3·24	3·31
20	1·87	2·10	2·26	2·38	2·48	2·56	2·63	2·47	2·73	2·91	3·04	3·14	3·22	3·29
24	1·84	2·07	2·23	2·35	2·44	2·52	2·59	2·43	2·68	2·85	2·97	3·07	3·15	3·22
30	1·82	2·04	2·20	2·31	2·40	2·48	2·55	2·38	2·62	2·79	2·91	3·01	3·08	3·15
40	1·80	2·02	2·17	2·28	2·37	2·44	2·51	2·34	2·57	2·73	2·85	2·94	3·02	3·08
60	1·78	1·99	2·14	2·25	2·33	2·41	2·47	2·30	2·52	2·68	2·79	2·88	2·95	3·01
120	1·76	1·97	2·11	2·21	2·30	2·37	2·43	2·25	2·48	2·62	2·73	2·82	2·89	2·95
∞	1·74	1·94	2·08	2·18	2·27	2·33	2·39	2·22	2·43	2·57	2·68	2·76	2·83	2·88

REFERENCES

GODWIN, H. J. (1945). On the distribution of the estimate of mean deviation obtained from samples from a normal population. *Biometrika*, **33**, 254.

HARTLEY, H. O. (1944). Studentization, or the elimination of the standard deviation of the parent population from the random sample distribution of statistics. *Biometrika*, **33**, 173.

HARTLEY, H. O. (1945). Note on the calculation of the distribution of the estimate of mean deviation in normal samples. *Biometrika*, **33**, 257.

IRWIN, J. O. (1925). On a criterion for the rejection of outlying observations. *Biometrika*, **17**, 238.

JONES, A. E. (1946). A useful method for the routine estimation of dispersion from large samples. *Biometrika*, **33**, 274.

MCKAY, A. T. (1935). The distribution of the difference between the extreme observation and the sample mean in samples of n from a normal universe. *Biometrika*, **27**, 466.

MCKAY, A. T. & PEARSON, E. S. (1933). A note on the distribution of range in samples of n. *Biometrika*, **25**, 415.

NAIR, K. R. (1947). A note on the mean deviation from the median. *Biometrika*, **34**, 360.

PEARSON, E. S. & CHANDRASEKAR, C. (1936). The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, **28**, 308.

PEARSON, E. S. & HARTLEY, H. O. (1943). Tables of the probability integral of the studentized range. *Biometrika*, **33**, 89.

'STUDENT' (1927). Errors of routine analysis. *Biometrika*, **19**, 151.

THOMPSON, W. R. (1935). On a criterion for the rejection of observations and the distribution of the ratio of deviation to sampling standard deviation. *Ann. Math. Statist.* **6**, 214.

TIPPETT, L. H. C. (1925). On the extreme individuals and the range of samples taken from a normal population. *Biometrika*, **17**, 364.

THE FISHER-YATES TEST OF SIGNIFICANCE IN 2×2 CONTINGENCY TABLES

By D. J. FINNEY

Lecturer in the Design and Analysis of Scientific Experiment, University of Oxford

1. Introduction

One of the tests of significance most frequently required in applications of statistics to biological problems is that for the 2×2 contingency table. Briefly, the problem may be typified by the following statement: Given two series of observations, with individual results assumed independent of one another and classified as either 'success' or 'failure', do the proportions of successes observed in the two series differ more widely than might reasonably be expected if the population values of these proportions are equal? If the expected numbers of successes and failures, calculated from the null hypothesis, are all moderately large, a simple and well-known form of χ^2 test may be used; this test gives a good approximation to the true probability of a deviation from equality as great as that observed, especially if the adjustment usually termed 'Yates's correction' is applied (Yates, 1934; Fisher, 1946, §§ 21, 21·01).

Not infrequently, however, the expected numbers are too small for the χ^2 test to be trusted. Fisher has recommended that, if any one of the expectations is less than 5, special steps should be taken; Aitken (1944) and Cramér (1946) suggest a minimum of 10. Investigations by Cochran (1942) indicate that no simple rule of this kind is completely adequate, and that the magnitudes of all four expectations affect the quality of the approximation. Yates (1934) and Fisher (1946, § 21·02) have given an exact method for the evaluation of the required probability, for use when small frequencies make any χ^2 approximation unreliable. Their method uses the four marginal totals as ancillary statistics: that is to say, the probability is obtained as a relative frequency amongst all configurations having the same marginal totals. More recent writers, notably Barnard (1947) and Pearson (1947), have questioned whether this assumption of fixed marginal totals is logically justifiable for all 2×2 contingency tables; Barnard has discussed the specifications of several distinct problems, and maintains that the Fisher-Yates test is applicable to only one of these. Nevertheless, the test is undoubtedly the right one for at least one important class of problem, and for this may be described as an exact method.†

The Fisher-Yates method requires rather tedious calculations, though the labour can be much reduced by use of a simple table of the logarithm of the factorial function, such as that given by Fisher & Yates (1948, Table XXX). The purpose of the present paper is to present a table of significance levels for their exact test which is applicable when all the frequencies are small. Yates (1934) and Fisher & Yates (1948, Table VIII) give a concise table which, with the aid of a small amount of calculation, leads to a good approximation to the exact probability; the new table provides tests of significance only, but has the advantage of direct entry instead of preliminary calculation. Rules for the use of the table are given in § 2, in sufficient detail for the non-mathematical reader; a brief account of the method of construction follows in § 3.

† The probability associated with any of the classes of problem distinguished by Barnard is satisfactorily approximated by that obtained from χ^2 when all expected frequencies are large.

2. The table of significance levels and its use

The table which follows may be used to test the significance of the deviation from proportionality in any 2 × 2 contingency table having both frequencies in one of its margins less than or equal to 15. The contingency table must first be put in the form

	Number of		Number of observations
	Successes	Failures	
Series I	a	$A-a$	A
Series II	b	$B-b$	B
Total	$a+b$	$A+B-a-b$	$A+B$

where Series I is defined to be that which makes $A \geqslant B$, and the type of observation conventionally regarded as a 'success' is that which makes $a/A \geqslant b/B$. For small integers, the right arrangement can usually be made on sight, especially if the second condition is used in the form

$$aB \geqslant bA.$$

Providing that $A \leqslant 15$, the table at the end of this section may then be entered in the section for A, the sub-section for B, and the line for a. If b is equal to or less than the integer in the column headed 0·05 or 0·01, a/A is significantly *greater than* b/B (single-tail test) at the probability level 0·05 or 0·01 respectively. If b is equal to or less than the integer in the column headed 0·025 or 0·005, a/A is significantly *different from* b/B (two-tail test) at the probability level 0·05 or 0·01 respectively. A dash, or absence of any entry, for some combination of A, B, and a indicates that no contingency table in that class is significant. The probability corresponding to b will generally be less than that shown at the head of the column, and the true numerical values are shown in small type. Those in the 0·025 and 0·005 columns should be doubled if used in a two-tail test.

From the mathematical point of view, the two pairs of marginal totals play equivalent roles in the test. Consequently, if only one set satisfies the condition that neither total exceeds 15, that margin may be conventionally regarded as relating to the two series, and the other classification must then be taken as 'success' and 'failure' (see example below). If both margins satisfy the condition, the two possible arrangements will lead to identical conclusions.

The use of the table presented here may be extended by noting that any contingency table more extreme than one known to show significant deviation from proportionality must itself be non-significant. Thus, even though one of each pair of marginal totals exceeds 15, a test of significance without calculation may still be possible. In particular, if the standard arrangement of the contingency table gives

a b	$A-a$ $B-b$	A B
$a+b$	$A+B-a-b$	$A+B$

with $A > 15 \geqslant B$ and $aB \geqslant bA$, the deviation from proportionality will be significant if

$15+a-A$ b	$A-a$ $B-b$	15 B
$15+a+b-A$	$A+B-a-b$	$15+B$

is significant (but not necessarily non-significant otherwise), and the deviation will be non-significant if

a b	$15-a$ $B-b$	15 B
$a+b$	$15+B-a-b$	$15+B$

is not significant (but not necessarily significant otherwise).

As an example of the use of the table, Lange's data on criminality among twin brothers or sisters of criminals (Fisher, 1946, § 21·01) may be examined. The contingency table below shows the numbers of twin brothers or sisters of criminals who had also been convicted, separately for monzygotic and dizygotic (but like-sexed) twins.

	Not convicted	Convicted	Total
Dizygotic	15	2	17
Monozygotic	3	10	13
Total	18	12	30

Since $15/17 > 3/13$ the category 'not convicted' is regarded as success. Consider the contingency table

13 3	2 10	15 13
16	12	28

in which $A = 15$, $B = 13$, and $a = 13$. The null hypothesis is that, in the general population, freedom from conviction is equally frequent amongst dizygotic and monozygotic sibs. If the only deviation from the null hypothesis which the investigator is prepared to consider is that monozygotic twins behave more similarly than dizygotic, he will require a one-tail significance test. The table below shows 4 as the 0·01 significance level of b (with a true probability of 0·004, not 0·010). Hence $b = 3$ is significant: by the rule given, the deviation from proportionality in the original table is significant evidence that criminality is more frequent among monozygotic twins of criminals than among dizygotic twins of criminals. Fisher gives the exact probability for the original data as 1/2150. On the other hand, if the investigator were concerned only to demonstrate a significant difference between the frequencies of freedom from conviction for the two types of twin, irrespective of whether monozygotic or dizygotic should show the higher value, he would use a two-tail test; for

$A = 15$, $B = 13$, $a = 13$, the value of b in the 0·005 column is again 4 (though generally it will be less than for the one-tail test), and again the evidence against the null hypothesis is judged significant, both for the modified and for the original contingency table.

As an illustration of the interchangeability of marginal totals, the same data may be examined by arranging the contingency table as:

	Dizygotic	Monozygotic	Total
Convicted	15	3	18
Not convicted	2	10	12
Total	17	13	30

in which 'dizygotic' is classed as success. Logically the meaning may be different, but the test of significance is the same. The modified contingency table

12	3	15
2	10	12
14	13	27

has $A = 15$, $B = 12$, $a = 12$; $b = 2$ is judged significant by comparison with the tabular values 3 at 0·01 or 2 at 0·005, and therefore again significance is attained on either the one-tail or the two-tail test.

The table follows on pp. 149–54.

3. Construction of the table

Any 2×2 contingency table to which a test of significance is to be applied must first be written in the standard form shown at the beginning of § 2, in which the event regarded as a success and the arrangement of the rows are so chosen that $A \geqslant B$ and $a/A \geqslant b/B$. If $a/A = b/B$, there is no need to proceed further, as the data then accord perfectly with the null hypothesis; if $a/A > b/B$, of necessity $a > b$. On the null hypothesis that the population values for the proportions of successes are equal, and for the specified set of marginal totals, the probability of this set of frequencies is (Fisher, 1946, § 21·02):

$$P_b = P(b \mid A, B, a+b)$$
$$= \frac{A!\,B!\,(a+b)!\,(A+B-a-b)!}{(A+B)!} \times \frac{1}{a!\,b!\,(A-a)!\,(B-b)!}.$$

The first factor in this expression is dependent only on the marginal totals, the second depends upon the internal cell frequencies. The probability that a deviation from equality of the two proportions as great as or greater than that observed should occur by chance is then P_b^*, where

$$P_b^* = P^*(b \mid A, B, a+b)$$
$$= P_b + P_{b-1} + P_{b-2} + \ldots + P_k,$$

and k is the greater of the two quantities 0 and $(a+b-A)$. In the summation the four marginal totals are kept fixed. Barnard (1947) has discussed the logic of the test in detail. He distinguishes certain classes of problem for which, in his view, the method described here is inappropriate; this topic has been a source of some controversy, to which the present paper is not intended as a contribution.

Table of significance levels of b

(Values of b in bold type; corresponding probabilities, P_b^*, in small type)

	a	Probability			
		0·05	0·025	0·01	0·005
A=3 B=3	3	**0** ·050	—	—	—
A=4 B=4	4	**0** ·014	**0** ·014	—	—
3	4	**0** ·029	—	—	—
A=5 B=5	5	**1** ·024	**1** ·024	**0** ·004	**0** ·004
	4	**0** ·024	**0** ·024	—	—
4	5	**1** ·048	**0** ·008	**0** ·008	—
	4	**0** ·040	—	—	—
3	5	**0** ·018	**0** ·018	—	—
2	5	**0** ·048	—	—	—
A=6 B=6	6	**2** ·030	**1** ·008	**1** ·008	**0** ·001
	5	**1** ·040	**0** ·008	**0** ·008	—
	4	**0** ·030	—	—	—
5	6	**1** ·015$^+$	**0** ·015$^+$	**0** ·002	**0** ·002
	5	**0** ·013	**0** ·013	—	—
	4	**0** ·045$^+$	—	—	—
4	6	**1** ·033	**0** ·005$^-$	**0** ·005$^-$	**0** ·005$^-$
	5	**0** ·024	**0** ·024	—	—
3	6	**0** ·012	**0** ·012	—	—
	5	**0** ·048	—	—	—
2	6	**0** ·036	—	—	—
A=7 B=7	7	**3** ·035$^-$	**2** ·010$^+$	**1** ·002	**1** ·002
	6	**1** ·015$^-$	**1** ·015$^-$	**0** ·002	**0** ·002
	5	**0** ·010$^+$	**0** ·010$^+$	—	—
	4	**0** ·035$^-$	—	—	—
6	7	**2** ·021	**2** ·021	**1** ·005$^-$	**1** ·005$^-$
	6	**1** ·025$^+$	**0** ·004	**0** ·004	**0** ·004
	5	**0** ·016	**0** ·016	—	—
	4	**0** ·049	—	—	—
5	7	**2** ·045$^+$	**1** ·010$^+$	**0** ·001	**0** ·001
	6	**1** ·045$^+$	**0** ·008	**0** ·008	—
	5	**0** ·027	—	—	—
4	7	**1** ·024	**1** ·024	**0** ·003	**0** ·003
	6	**0** ·015$^+$	**0** ·015$^+$	—	—
	5	**0** ·045$^+$	—	—	—
3	7	**0** ·008	**0** ·008	**0** ·008	—
	6	**0** ·033	—	—	—
2	7	**0** ·028	—	—	—

	a	Probability			
		0·05	0·025	0·01	0·005
A=8 B=8	8	**4** ·038	**3** ·013	**2** ·003	**2** ·003
	7	**2** ·020	**2** ·020	**1** ·005$^+$	**0** ·001
	6	**1** ·020	**1** ·020	**0** ·003	**0** ·003
	5	**0** ·013	**0** ·013	—	—
	4	**0** ·038	—	—	—
7	8	**3** ·026	**2** ·007	**2** ·007	**1** ·001
	7	**2** ·035$^-$	**1** ·009	**1** ·009	**0** ·001
	6	**1** ·032	**0** ·006	**0** ·006	—
	5	**0** ·019	**0** ·019	—	—
6	8	**2** ·015$^-$	**2** ·015$^-$	**1** ·003	**1** ·003
	7	**1** ·016	**1** ·016	**0** ·002	**0** ·002
	6	**0** ·009	**0** ·009	**0** ·009	—
	5	**0** ·028	—	—	—
5	8	**2** ·035$^-$	**1** ·007	**1** ·007	**0** ·001
	7	**1** ·032	**0** ·005$^-$	**0** ·005$^-$	**0** ·005$^-$
	6	**0** ·016	**0** ·016	—	—
	5	**0** ·044	—	—	—
4	8	**1** ·018	**1** ·018	**0** ·002	**0** ·002
	7	**0** ·010$^+$	**0** ·010$^+$	—	—
	6	**0** ·030	—	—	—
3	8	**0** ·006	**0** ·006	**0** ·006	—
	7	**0** ·024	**0** ·024	—	—
2	8	**0** ·022	**0** ·022	—	—
A=9 B=9	9	**5** ·041	**4** ·015$^-$	**3** ·005$^-$	**3** ·005$^-$
	8	**3** ·025$^-$	**3** ·025$^-$	**2** ·008	**1** ·002
	7	**2** ·028	**1** ·008	**1** ·008	**0** ·001
	6	**1** ·025$^-$	**1** ·025$^-$	**0** ·005$^-$	**0** ·005$^-$
	5	**0** ·015$^-$	**0** ·015$^-$	—	—
	4	**0** ·041	—	—	—
8	9	**4** ·029	**3** ·009	**3** ·009	**2** ·002
	8	**3** ·043	**2** ·013	**1** ·003	**1** ·003
	7	**2** ·044	**1** ·012	**0** ·002	**0** ·002
	6	**1** ·036	**0** ·007	**0** ·007	—
	5	**0** ·020	**0** ·020	—	—
7	9	**3** ·019	**3** ·019	**2** ·005$^-$	**2** ·005$^-$
	8	**2** ·024	**2** ·024	**1** ·006	**0** ·001
	7	**1** ·020	**1** ·020	**0** ·003	**0** ·003
	6	**0** ·010$^+$	**0** ·010$^+$	—	—
	5	**0** ·029	—	—	—
6	9	**3** ·044	**2** ·011	**1** ·002	**1** ·002
	8	**2** ·047	**1** ·011	**0** ·001	**0** ·001
	7	**1** ·035$^-$	**0** ·006	**0** ·006	—
	6	**0** ·017	**0** ·017	—	—
	5	**0** ·042	—	—	—

Table of significance levels of b (continued)

	a	Probability			
		0·05	0·025	0·01	0·005
A=9 B=5	9	2 ·027	1 ·005⁻	1 ·005⁻	1 ·005⁻
	8	1 ·023	1 ·023	0 ·003	0 ·003
	7	0 ·010⁺	0 ·010⁺	—	—
	6	0 ·028	—	—	—
4	9	1 ·014	1 ·014	0 ·001	0 ·001
	8	0 ·007	0 ·007	0 ·007	—
	7	0 ·021	0 ·021	—	—
	6	0 ·049	—	—	—
3	9	1 ·045⁺	0 ·005⁻	0 ·005⁻	0 ·005⁻
	8	0 ·018	0 ·018	—	—
	7	0 ·045⁺	—	—	—
2	9	0 ·018	0 ·018	—	—
A=10 B=10	10	6 ·043	5 ·016	4 ·005⁺	3 ·002
	9	4 ·029	3 ·010⁻	3 ·010⁻	2 ·003
	8	3 ·035⁻	2 ·012	1 ·003	1 ·003
	7	2 ·035⁻	1 ·010⁻	1 ·010⁻	0 ·002
	6	1 ·029	0 ·005⁺	0 ·005⁺	—
	5	0 ·016	0 ·016	—	—
	4	0 ·043	—	—	—
9	10	5 ·033	4 ·011	3 ·003	3 ·003
	9	4 ·050⁻	3 ·017	2 ·005⁻	2 ·005⁻
	8	2 ·019	2 ·019	1 ·004	1 ·004
	7	1 ·015⁻	1 ·015⁻	0 ·002	0 ·002
	6	1 ·040	0 ·008	0 ·008	—
	5	0 ·022	0 ·022	—	—
8	10	4 ·023	4 ·023	3 ·007	2 ·002
	9	3 ·032	2 ·009	2 ·009	1 ·002
	8	2 ·031	1 ·008	1 ·008	0 ·001
	7	1 ·023	1 ·023	0 ·004	0 ·004
	6	0 ·011	0 ·011	—	—
	5	0 ·029	—	—	—
7	10	3 ·015⁻	3 ·015⁻	2 ·003	2 ·003
	9	2 ·018	2 ·018	1 ·004	1 ·004
	8	1 ·013	1 ·013	0 ·002	0 ·002
	7	1 ·036	0 ·006	0 ·006	—
	6	0 ·017	0 ·017	—	—
	5	0 ·041	—	—	—
6	10	3 ·036	2 ·008	2 ·008	1 ·001
	9	2 ·036	1 ·008	1 ·008	0 ·001
	8	1 ·024	1 ·024	0 ·003	0 ·003
	7	0 ·010⁺	0 ·010⁻	—	—
	6	0 ·026	—	—	—
5	10	2 ·022	2 ·022	1 ·004	1 ·004
	9	1 ·017	1 ·017	0 ·002	0 ·002
	8	1 ·047	0 ·007	0 ·007	—
	7	0 ·019	0 ·019	—	—
	6	0 ·042	—	—	—

	a	Probability			
		0·05	0·025	0·01	0·005
A=10 B=4	10	1 ·011	1 ·011	0 ·001	0 ·001
	9	1 ·041	0 ·005⁻	0 ·005⁻	0 ·005⁻
	8	0 ·015⁻	0 ·015⁻	—	—
	7	0 ·035⁻	—	—	—
3	10	1 ·038	0 ·003	0 ·003	0 ·003
	9	0 ·014	0 ·014	—	—
	8	0 ·035⁻	—	—	—
2	10	0 ·015⁺	0 ·015⁺	—	—
	9	0 ·045⁺	—	—	—
A=11 B=11	11	7 ·045⁺	6 ·018	5 ·006	4 ·002
	10	5 ·032	4 ·012	3 ·004	3 ·004
	9	4 ·040	3 ·015⁻	2 ·004	2 ·004
	8	3 ·043	2 ·015⁻	1 ·004	1 ·004
	7	2 ·040	1 ·012	0 ·002	0 ·002
	6	1 ·032	0 ·006	0 ·006	—
	5	0 ·018	0 ·018	—	—
	4	0 ·045⁺	—	—	—
10	11	6 ·035⁺	5 ·012	4 ·004	4 ·004
	10	4 ·021	4 ·021	3 ·007	2 ·002
	9	3 ·024	3 ·024	2 ·007	1 ·002
	8	2 ·023	2 ·023	1 ·006	0 ·001
	7	1 ·017	1 ·017	0 ·003	0 ·003
	6	1 ·043	0 ·009	0 ·009	—
	5	0 ·023	0 ·023	—	—
9	11	5 ·026	4 ·008	4 ·008	3 ·002
	10	4 ·038	3 ·012	2 ·003	2 ·003
	9	3 ·040	2 ·012	1 ·003	1 ·003
	8	2 ·035	1 ·009	1 ·009	0 ·001
	7	1 ·025	1 ·025	0 ·004	0 ·004
	6	0 ·012	0 ·012	—	—
	5	0 ·030	—	—	—
8	11	4 ·018	4 ·018	3 ·005⁻	3 ·005⁻
	10	3 ·024	3 ·024	2 ·006	1 ·001
	9	2 ·022	2 ·022	1 ·005	1 ·005
	8	1 ·015⁻	1 ·015⁻	0 ·002	0 ·002
	7	1 ·037	0 ·007	0 ·007	—
	6	0 ·017	0 ·017	—	—
	5	0 ·040	—	—	—
7	11	4 ·043	3 ·011	2 ·002	2 ·002
	10	3 ·047	2 ·013	1 ·002	1 ·002
	9	2 ·039	1 ·009	1 ·009	0 ·001
	8	1 ·025⁻	1 ·025⁻	0 ·004	0 ·004
	7	0 ·010⁺	0 ·010⁺	—	—
	6	0 ·025	0 ·025⁻	—	—
6	11	3 ·029	2 ·006	2 ·006	1 ·001
	10	2 ·028	1 ·005⁺	1 ·005⁺	0 ·001
	9	1 ·018	1 ·018	0 ·002	0 ·002

Table of significance levels of b (continued)

	a	Probability			
		0·05	0·025	0·01	0·005
A=11 B=6	8	1 ·043	0 ·007	0 ·007	—
	7	0 ·017	0 ·017	—	—
	6	0 ·037	—	—	—
5	11	2 ·018	2 ·018	1 ·003	1 ·003
	10	1 ·013	1 ·013	0 ·001	0 ·001
	9	1 ·036	0 ·005⁻	0 ·005⁻	0 ·005⁻
	8	0 ·013	0 ·013	—	—
	7	0 ·029	—	—	—
4	11	1 ·009	1 ·009	1 ·009	0 ·001
	10	1 ·033	0 ·004	0 ·004	0 ·004
	9	0 ·011	0 ·011	—	—
	8	0 ·026	—	—	—
3	11	1 ·033	0 ·003	0 ·003	0 ·003
	10	0 ·011	0 ·011	—	—
	9	0 ·027	—	—	—
2	11	0 ·013	0 ·013	—	—
	10	0 ·038	—	—	—
A=12 B=12	12	8 ·047	7 ·019	6 ·007	5 ·002
	11	6 ·034	5 ·014	4 ·005⁻	4 ·005⁻
	10	5 ·045⁻	4 ·018	3 ·006	2 ·002
	9	4 ·050⁻	3 ·020	2 ·006	1 ·001
	8	3 ·050⁻	2 ·018	1 ·005⁻	1 ·005⁻
	7	2 ·045⁻	1 ·014	0 ·002	0 ·002
	6	1 ·034	0 ·007	0 ·007	—
	5	0 ·019	0 ·019	—	—
	4	0 ·047	—	—	—
11	12	7 ·037	6 ·014	5 ·005⁻	5 ·005⁻
	11	5 ·024	5 ·024	4 ·008	3 ·002
	10	4 ·029	3 ·010⁺	2 ·003	2 ·003
	9	3 ·030	2 ·009	2 ·009	1 ·002
	8	2 ·026	1 ·007	1 ·007	0 ·001
	7	1 ·019	1 ·019	0 ·003	0 ·003
	6	1 ·045⁻	0 ·009	0 ·009	—
	5	0 ·024	0 ·024	—	—
10	12	6 ·029	5 ·010⁻	5 ·010⁻	4 ·003
	11	5 ·043	4 ·015⁺	3 ·005⁻	3 ·005⁻
	10	4 ·048	3 ·017	2 ·005⁻	2 ·005⁻
	9	3 ·046	2 ·015⁻	1 ·004	1 ·004
	8	2 ·038	1 ·010⁺	0 ·002	0 ·002
	7	1 ·026	0 ·005⁻	0 ·005⁻	0 ·005⁻
	6	0 ·012	0 ·012	—	—
	5	0 ·030	—	—	—
9	12	5 ·021	5 ·021	4 ·006	3 ·002
	11	4 ·029	3 ·009	3 ·009	2 ·002
	10	3 ·029	2 ·008	2 ·008	1 ·002
	9	2 ·024	2 ·024	1 ·006	0 ·001
	8	1 ·016	1 ·016	0 ·002	0 ·002

	a	Probability			
		0·05	0·025	0·01	0·005
A=12 B=9	7	1 ·037	0 ·007	0 ·007	—
	6	0 ·017	0 ·017	—	—
	5	0 ·039	—	—	—
8	12	5 ·049	4 ·014	3 ·004	3 ·004
	11	3 ·018	3 ·018	2 ·004	2 ·004
	10	2 ·015⁺	2 ·015⁺	1 ·003	1 ·003
	9	2 ·040	1 ·010⁻	1 ·010⁻	0 ·001
	8	1 ·025⁻	1 ·025⁻	0 ·004	0 ·004
	7	0 ·010⁺	0 ·010⁺	—	—
	6	0 ·024	0 ·024	—	—
7	12	4 ·036	3 ·009	3 ·009	2 ·002
	11	3 ·038	2 ·010⁻	2 ·010⁻	1 ·002
	10	2 ·029	1 ·006	1 ·006	0 ·001
	9	1 ·017	1 ·017	0 ·002	0 ·002
	8	1 ·040	0 ·007	0 ·007	—
	7	0 ·016	0 ·016	—	—
	6	0 ·034	—	—	—
6	12	3 ·025⁻	3 ·025⁻	2 ·005⁻	2 ·005⁻
	11	2 ·022	2 ·022	1 ·004	1 ·004
	10	1 ·013	1 ·013	0 ·002	0 ·002
	9	1 ·032	0 ·005⁻	0 ·005⁻	0 ·005⁻
	8	0 ·011	0 ·011	—	—
	7	0 ·025⁻	0 ·025⁻	—	—
	6	0 ·050⁻	—	—	—
5	12	2 ·015⁻	2 ·015⁻	1 ·002	1 ·002
	11	1 ·010⁻	1 ·010⁻	1 ·010⁻	0 ·001
	10	1 ·028	0 ·003	0 ·003	0 ·003
	9	0 ·009	0 ·009	0 ·009	—
	8	0 ·020	0 ·020	—	—
	7	0 ·041	—	—	—
4	12	2 ·050	1 ·007	1 ·007	0 ·001
	11	1 ·027	0 ·003	0 ·003	0 ·003
	10	0 ·008	0 ·008	0 ·008	—
	9	0 ·019	0 ·019	—	—
	8	0 ·038	—	—	—
3	12	1 ·029	0 ·002	0 ·002	0 ·002
	11	0 ·009	0 ·009	0 ·009	—
	10	0 ·022	0 ·022	—	—
	9	0 ·044	—	—	—
2	12	0 ·011	0 ·011	—	—
	11	0 ·033	—	—	—
A=13 B=13	13	9 ·048	8 ·020	7 ·007	6 ·003
	12	7 ·037	6 ·015⁺	5 ·006	4 ·002
	11	6 ·048	5 ·021	4 ·008	3 ·002
	10	4 ·024	4 ·024	3 ·008	2 ·002
	9	3 ·024	3 ·024	2 ·008	1 ·002
	8	2 ·021	2 ·021	1 ·006	0 ·001

Table of significance levels of b (continued)

	a	Probability			
		0·05	0·025	0·01	0·005
A=13 B=13	7	2 ·048	1 ·015⁺	0 ·003	0 ·003
	6	1 ·037	0 ·007	0 ·007	—
	5	0 ·020	0 ·020	—	—
	4	0 ·048	—	—	—
12	13	8 ·039	7 ·015⁻	6 ·005⁺	5 ·002
	12	6 ·027	5 ·010⁻	5 ·010⁻	4 ·003
	11	5 ·033	4 ·013	3 ·004	3 ·004
	10	4 ·036	3 ·013	2 ·004	2 ·004
	9	3 ·034	2 ·011	1 ·003	1 ·003
	8	2 ·029	1 ·008	1 ·008	0 ·001
	7	1 ·020	1 ·020	0 ·004	0 ·004
	6	1 ·046	0 ·010⁻	0 ·010⁻	—
	5	0 ·024	0 ·024	—	—
11	13	7 ·031	6 ·011	5 ·003	5 ·003
	12	6 ·048	5 ·018	4 ·006	3 ·002
	11	4 ·021	4 ·021	3 ·007	2 ·002
	10	3 ·021	3 ·021	2 ·006	1 ·001
	9	3 ·050⁻	2 ·017	1 ·004	1 ·004
	8	2 ·040	1 ·011	0 ·002	0 ·002
	7	1 ·027	0 ·005⁻	0 ·005⁻	0 ·005⁻
	6	0 ·013	0 ·013	—	—
	5	0 ·030	—	—	—
10	13	6 ·024	6 ·024	5 ·007	4 ·002
	12	5 ·035⁻	4 ·012	3 ·003	3 ·003
	11	4 ·037	3 ·012	2 ·003	2 ·003
	10	3 ·033	2 ·010⁺	1 ·002	1 ·002
	9	2 ·026	1 ·006	1 ·006	0 ·001
	8	1 ·017	1 ·017	0 ·003	0 ·003
	7	1 ·038	0 ·007	0 ·007	—
	6	0 ·017	0 ·017	—	—
	5	0 ·038	—	—	—
9	13	5 ·017	5 ·017	4 ·005⁻	4 ·005⁻
	12	4 ·023	4 ·023	3 ·007	2 ·001
	11	3 ·022	3 ·022	2 ·006	1 ·001
	10	2 ·017	2 ·017	1 ·004	1 ·004
	9	2 ·040	1 ·010⁺	0 ·001	0 ·001
	8	1 ·025⁻	1 ·025⁻	0 ·004	0 ·004
	7	0 ·010⁺	0 ·010⁺	—	—
	6	0 ·023	0 ·023	—	—
	5	0 ·049	—	—	—
8	13	5 ·042	4 ·012	3 ·003	3 ·003
	12	4 ·047	3 ·014	2 ·003	2 ·003
	11	3 ·041	2 ·011	1 ·002	1 ·002
	10	2 ·029	1 ·007	1 ·007	0 ·001
	9	1 ·017	1 ·017	0 ·002	0 ·002
	8	1 ·037	0 ·006	0 ·006	—
	7	0 ·015⁻	0 ·015⁻	—	—
	6	0 ·032	—	—	—
7	13	4 ·031	3 ·007	3 ·007	2 ·001
	12	3 ·031	2 ·007	2 ·007	1 ·001

	a	Probability			
		0·05	0·025	0·01	0·005
A=13 B=7	11	2 ·022	2 ·022	1 ·004	1 ·004
	10	1 ·012	1 ·012	0 ·002	0 ·002
	9	1 ·029	0 ·004	0 ·004	0 ·004
	8	0 ·010⁺	0 ·010⁺	—	—
	7	0 ·022	0 ·022	—	—
	6	0 ·044	—	—	—
6	13	3 ·021	3 ·021	2 ·004	2 ·004
	12	2 ·017	2 ·017	1 ·003	1 ·003
	11	2 ·046	1 ·010⁻	1 ·010⁻	0 ·001
	10	1 ·024	1 ·024	0 ·003	0 ·003
	9	1 ·050⁻	0 ·008	0 ·008	—
	8	0 ·017	0 ·017	—	—
	7	0 ·034	—	—	—
5	13	2 ·012	2 ·012	1 ·002	1 ·002
	12	2 ·044	1 ·008	1 ·008	0 ·001
	11	1 ·022	1 ·022	0 ·002	0 ·002
	10	1 ·047	0 ·007	0 ·007	—
	9	0 ·015⁻	0 ·015⁻	—	—
	8	0 ·029	—	—	—
4	13	2 ·044	1 ·006	1 ·006	0 ·000
	12	1 ·022	1 ·022	0 ·002	0 ·002
	11	0 ·006	0 ·006	0 ·006	—
	10	0 ·015⁻	0 ·015⁻	—	—
	9	0 ·029	—	—	—
3	13	1 ·025	1 ·025	0 ·002	0 ·002
	12	0 ·007	0 ·007	0 ·007	—
	11	0 ·018	0 ·018	—	—
	10	0 ·036	—	—	—
2	13	0 ·010⁻	0 ·010⁻	0 ·010⁻	—
	12	0 ·029	—	—	—
A=14 B=14	14	10 ·049	9 ·020	8 ·008	7 ·003
	13	8 ·038	7 ·016	6 ·006	5 ·002
	12	6 ·023	6 ·023	5 ·009	4 ·003
	11	5 ·027	4 ·011	3 ·004	3 ·004
	10	4 ·028	3 ·011	2 ·003	2 ·003
	9	3 ·027	2 ·009	2 ·009	1 ·002
	8	2 ·023	2 ·023	1 ·006	0 ·001
	7	1 ·016	1 ·016	0 ·003	0 ·003
	6	1 ·038	0 ·008	0 ·008	—
	5	0 ·020	0 ·020	—	—
	4	0 ·049	—	—	—
13	14	9 ·041	8 ·016	7 ·006	6 ·002
	13	7 ·029	6 ·011	5 ·004	5 ·004
	12	6 ·037	5 ·015⁺	4 ·005⁺	3 ·002
	11	5 ·041	4 ·017	3 ·006	2 ·001
	10	4 ·041	3 ·016	2 ·005⁻	2 ·005⁻
	9	3 ·038	2 ·013	1 ·003	1 ·003
	8	2 ·031	1 ·009	1 ·009	0 ·001

Table of significance levels of b (continued)

	a	Probability			
		0·05	0·025	0·01	0·005
A=14 B=13	7	1 ·021	1 ·021	0 ·004	0 ·004
	6	1 ·048	0 ·010+	—	—
	5	0 ·025−	0 ·025−	—	—
12	14	8 ·033	7 ·012	6 ·004	6 ·004
	13	6 ·021	6 ·021	5 ·007	4 ·002
	12	5 ·025+	4 ·009	4 ·009	3 ·003
	11	4 ·026	3 ·009	3 ·009	2 ·002
	10	3 ·024	3 ·024	2 ·007	1 ·002
	9	2 ·019	2 ·019	1 ·005−	1 ·005−
	8	2 ·042	1 ·012	0 ·002	0 ·002
	7	1 ·028	0 ·005+	0 ·005+	—
	6	0 ·013	0 ·013	—	—
	5	0 ·030	—	—	—
11	14	7 ·026	6 ·009	6 ·009	5 ·003
	13	6 ·039	5 ·014	4 ·004	4 ·004
	12	5 ·043	4 ·016	3 ·005−	3 ·005−
	11	4 ·042	3 ·015−	2 ·004	2 ·004
	10	3 ·036	2 ·011	1 ·003	1 ·003
	9	2 ·027	1 ·007	1 ·007	0 ·001
	8	1 ·017	1 ·017	0 ·003	0 ·003
	7	1 ·038	0 ·007	0 ·007	—
	6	0 ·017	0 ·017	—	—
	5	0 ·038	—	—	—
10	14	6 ·020	6 ·020	5 ·006	4 ·002
	13	5 ·028	4 ·009	4 ·009	3 ·002
	12	4 ·028	3 ·009	3 ·009	2 ·002
	11	3 ·024	3 ·024	2 ·007	1 ·001
	10	2 ·018	2 ·018	1 ·004	1 ·004
	9	2 ·040	1 ·011	0 ·002	0 ·002
	8	1 ·024	1 ·024	0 ·004	0 ·004
	7	0 ·010−	0 ·010−	0 ·010−	—
	6	0 ·022	0 ·022	—	—
	5	0 ·047	—	—	—
9	14	6 ·047	5 ·014	4 ·004	4 ·004
	13	4 ·018	4 ·018	3 ·005−	3 ·005−
	12	3 ·017	3 ·017	2 ·004	2 ·004
	11	3 ·042	2 ·012	1 ·002	1 ·002
	10	2 ·029	1 ·007	1 ·007	0 ·001
	9	1 ·017	1 ·017	0 ·002	0 ·002
	8	1 ·036	0 ·006	0 ·006	—
	7	0 ·014	0 ·014	—	—
	6	0 ·030	—	—	—
8	14	5 ·036	4 ·010−	4 ·010−	3 ·002
	13	4 ·039	3 ·011	2 ·002	2 ·002
	12	3 ·032	2 ·008	2 ·008	1 ·001
	11	2 ·022	2 ·022	1 ·005−	1 ·005−
	10	2 ·048	1 ·012	0 ·002	0 ·002
	9	1 ·026	0 ·004	0 ·004	0 ·004
	8	0 ·009	0 ·009	0 ·009	—
	7	0 ·020	0 ·020	—	—
	6	0 ·040	—	—	—

	a	Probability			
		0·05	0·025	0·01	0·005
A=14 B=7	14	4 ·026	3 ·006	3 ·006	2 ·001
	13	3 ·025	2 ·006	2 ·006	1 ·001
	12	2 ·017	2 ·017	1 ·003	1 ·003
	11	2 ·041	1 ·009	1 ·009	0 ·001
	10	1 ·021	1 ·021	0 ·003	0 ·003
	9	1 ·043	0 ·007	0 ·007	—
	8	0 ·015−	0 ·015−	—	—
	7	0 ·030	—	—	—
6	14	3 ·018	3 ·018	2 ·003	2 ·003
	13	2 ·014	2 ·014	1 ·002	1 ·002
	12	2 ·037	1 ·007	1 ·007	0 ·001
	11	1 ·018	1 ·018	0 ·002	0 ·002
	10	1 ·038	0 ·005+	0 ·005+	—
	9	0 ·012	0 ·012	—	—
	8	0 ·024	0 ·024	—	—
	7	0 ·044	—	—	—
5	14	2 ·010+	2 ·010+	1 ·001	1 ·001
	13	2 ·037	1 ·006	1 ·006	0 ·001
	12	1 ·017	1 ·017	0 ·002	0 ·002
	11	1 ·038	0 ·005−	0 ·005−	0 ·005−
	10	0 ·011	0 ·011	—	—
	9	0 ·022	0 ·022	—	—
	8	0 ·040	—	—	—
4	14	2 ·039	1 ·005−	1 ·005−	1 ·005−
	13	1 ·019	1 ·019	0 ·002	0 ·002
	12	1 ·044	0 ·005−	0 ·005−	0 ·005−
	11	0 ·011	0 ·011	—	—
	10	0 ·023	0 ·023	—	—
	9	0 ·041	—	—	—
3	14	1 ·022	1 ·022	0 ·001	0 ·001
	13	0 ·006	0 ·006	0 ·006	—
	12	0 ·015−	0 ·015−	—	—
	11	0 ·029	—	—	—
2	14	0 ·008	0 ·008	0 ·008	—
	13	0 ·025	0 ·025	—	—
	12	0 ·050	—	—	—
A=15 B=15	15	11 ·050−	10 ·021	9 ·008	8 ·003
	14	9 ·040	8 ·018	7 ·007	6 ·003
	13	7 ·025+	6 ·010+	5 ·004	5 ·004
	12	6 ·030	5 ·013	4 ·005−	4 ·005−
	11	5 ·033	4 ·013	3 ·005−	3 ·005−
	10	4 ·033	3 ·013	2 ·004	2 ·004
	9	3 ·030	2 ·010+	1 ·003	1 ·003
	8	2 ·025+	1 ·007	1 ·007	0 ·001
	7	1 ·018	1 ·018	0 ·003	0 ·003
	6	1 ·040	0 ·008	0 ·008	—
	5	0 ·021	0 ·021	—	—
	4	0 ·050−	—	—	—

Table of significance levels of b (continued)

	a	Probability			
		0·05	0·025	0·01	0·005
A=15 B=14	15	10 ·042	9 ·017	8 ·006	7 ·002
	14	8 ·031	7 ·013	6 ·005⁻	6 ·005⁻
	13	7 ·041	6 ·017	5 ·007	4 ·002
	12	6 ·046	5 ·020	4 ·007	3 ·002
	11	5 ·048	4 ·020	3 ·007	2 ·002
	10	4 ·046	3 ·018	2 ·006	1 ·001
	9	3 ·041	2 ·014	1 ·004	1 ·004
	8	2 ·033	1 ·009	1 ·009	0 ·001
	7	1 ·022	1 ·022	0 ·004	0 ·004
	6	1 ·049	0 ·011	—	—
	5	0 ·025⁺	—	—	—
13	15	9 ·035⁻	8 ·013	7 ·005⁻	7 ·005⁻
	14	7 ·023	7 ·023	6 ·009	5 ·003
	13	6 ·029	5 ·011	4 ·004	4 ·004
	12	5 ·031	4 ·012	3 ·004	3 ·004
	11	4 ·030	3 ·011	2 ·003	2 ·003
	10	3 ·026	2 ·008	2 ·008	1 ·002
	9	2 ·020	2 ·020	1 ·005⁺	0 ·001
	8	2 ·043	1 ·013	0 ·002	0 ·002
	7	1 ·029	0 ·005⁺	0 ·005⁺	—
	6	0 ·013	0 ·013	—	—
	5	0 ·031	—	—	—
12	15	8 ·028	7 ·010⁻	7 ·010⁻	6 ·003
	14	7 ·043	6 ·016	5 ·006	4 ·002
	13	6 ·049	5 ·019	4 ·007	3 ·002
	12	5 ·049	4 ·019	3 ·006	2 ·002
	11	4 ·045⁺	3 ·017	2 ·005⁻	2 ·005⁻
	10	3 ·038	2 ·012	1 ·003	1 ·003
	9	2 ·028	1 ·007	1 ·007	0 ·001
	8	1 ·018	1 ·018	0 ·003	0 ·003
	7	1 ·038	0 ·007	0 ·007	—
	6	0 ·017	0 ·017	—	—
	5	0 ·037	—	—	—
11	15	7 ·022	7 ·022	6 ·007	5 ·002
	14	6 ·032	5 ·011	4 ·003	4 ·003
	13	5 ·034	4 ·012	3 ·003	3 ·003
	12	4 ·032	3 ·010⁺	2 ·003	2 ·003
	11	3 ·026	2 ·008	2 ·008	1 ·002
	10	2 ·019	2 ·019	1 ·004	1 ·004
	9	2 ·040	1 ·011	0 ·002	0 ·002
	8	1 ·024	1 ·024	0 ·004	0 ·004
	7	1 ·049	0 ·010⁻	0 ·010⁻	—
	6	0 ·022	0 ·022	—	—
	5	0 ·046	—	—	—
10	15	6 ·047	6 ·017	5 ·005⁻	5 ·005⁻
	14	5 ·023	5 ·023	4 ·007	3 ·002
	13	4 ·022	4 ·022	3 ·007	2 ·001
	12	3 ·018	3 ·018	2 ·005⁻	2 ·005⁻
	11	3 ·042	2 ·013	1 ·003	1 ·003
	10	2 ·029	1 ·007	1 ·007	0 ·001
	9	1 ·016	1 ·016	0 ·002	0 ·002
	8	1 ·034	0 ·006	0 ·006	—
	7	0 ·013	0 ·013	—	—
	6	0 ·028	—	—	—
9	15	6 ·042	5 ·012	4 ·003	4 ·003
	14	5 ·047	4 ·015⁻	3 ·004	3 ·004

	a	Probability			
		0·05	0·025	0·01	0·005
A=15 B=9	13	4 ·042	3 ·013	2 ·003	2 ·003
	12	3 ·032	2 ·009	2 ·009	1 ·002
	11	2 ·021	2 ·021	1 ·005⁻	1 ·005⁻
	10	2 ·045⁻	1 ·011	0 ·002	0 ·002
	9	1 ·024	1 ·024	0 ·004	0 ·004
	8	1 ·048	0 ·009	0 ·009	—
	7	0 ·019	0 ·019	—	—
	6	0 ·037	—	—	—
8	15	5 ·032	4 ·008	4 ·008	3 ·002
	14	4 ·033	3 ·009	3 ·009	2 ·002
	13	3 ·026	2 ·006	2 ·006	1 ·001
	12	2 ·017	2 ·017	1 ·003	1 ·003
	11	2 ·037	1 ·008	1 ·008	0 ·001
	10	1 ·019	1 ·019	0 ·003	0 ·003
	9	1 ·038	0 ·006	0 ·006	—
	8	0 ·013	0 ·013	—	—
	7	0 ·026	—	—	—
	6	0 ·050⁻	—	—	—
7	15	4 ·023	4 ·023	3 ·005⁻	3 ·005⁻
	14	3 ·021	3 ·021	2 ·004	2 ·004
	13	2 ·014	2 ·014	1 ·002	1 ·002
	12	2 ·032	1 ·007	1 ·007	0 ·001
	11	1 ·015⁺	1 ·015⁺	0 ·002	0 ·002
	10	1 ·032	0 ·005⁻	0 ·005⁻	0 ·005⁻
	9	0 ·010⁺	0 ·010⁺	—	—
	8	0 ·020	0 ·020	—	—
	7	0 ·038	—	—	—
6	15	3 ·015⁺	3 ·015⁺	2 ·003	2 ·003
	14	2 ·011	2 ·011	1 ·002	1 ·002
	13	2 ·031	1 ·006	1 ·006	0 ·001
	12	1 ·014	1 ·014	0 ·002	0 ·002
	11	1 ·029	0 ·004	0 ·004	0 ·004
	10	0 ·009	0 ·009	0 ·009	—
	9	0 ·017	0 ·017	—	—
	8	0 ·032	—	—	—
5	15	2 ·009	2 ·009	2 ·009	1 ·001
	14	2 ·032	1 ·005⁻	1 ·005⁻	1 ·005⁻
	13	1 ·014	1 ·014	0 ·001	0 ·001
	12	1 ·031	0 ·004	0 ·004	0 ·004
	11	0 ·008	0 ·008	0 ·008	—
	10	0 ·016	0 ·016	—	—
	9	0 ·030	—	—	—
4	15	2 ·035⁺	1 ·004	1 ·004	1 ·004
	14	1 ·016	1 ·016	0 ·001	0 ·001
	13	1 ·037	0 ·004	0 ·004	0 ·004
	12	0 ·009	0 ·009	0 ·009	—
	11	0 ·018	0 ·018	—	—
	10	0 ·033	—	—	—
3	15	1 ·020	1 ·020	0 ·001	0 ·001
	14	0 ·005⁻	0 ·005⁻	0 ·005⁻	0 ·005⁻
	13	0 ·012	0 ·012	—	—
	12	0 ·025⁻	0 ·025⁻	—	—
	11	0 ·043	—	—	—
2	15	0 ·007	0 ·007	0 ·007	—
	14	0 ·022	0 ·022	—	—
	13	0 ·044	—	—	—

The table given in § 2 enables tests of significance, at probability levels of 0·05, 0·025, 0·01 and 0·005, to be made by direct reference for any 2×2 contingency table having $B \leqslant A \leqslant 15$. The section of the table for any particular pair of values A, B was constructed by giving to $n = a+b$ successively all integral values from $(A+B)$ down to zero; for each n, b was given in turn the values $k, k+1, k+2, \ldots$ (where k is the greater of the two quantities 0 and $n-A$), P_b was formed and the cumulative sum P_b^* was recorded. The calculations with any set of A, B, n were stopped as soon as P_b^* exceeded 0·05 or would obviously exceed 0·05 for the next higher value of b. The calculations were tedious, but by no means as lengthy as might appear from this account, for after a little experience extreme values of n, which would always give $P_b^* > 0{\cdot}05$, could be ignored, and trends in P_b^* often enabled values of b which would exceed the limit to be foreseen without calculation of P_b. Of course, in any case of doubt, P_b^* was evaluated.

The probabilities were determined to five places of decimals, but any full publication of these would occupy a great deal of space. For many practical purposes, all that is required is a test of significance at an arbitrary level of probability. This can be conveniently made by use of a table showing, for each combination of A, B, and a, the greatest value of b for which P_b^* lies beyond the chosen significance level. The values of b required can be seen immediately from systematic inspection of the calculations just described. For example, the calculations show the following 2×2 configurations with their associated values of P_b^*.

8	3	11	8	3	11	8	3	11	8	3	11
3	7	10	2	8	10	1	9	10	0	10	10
11	10	21	10	11	21	9	12	21	8	13	21
$P_b^* = 0{\cdot}0635$			$P_b^* = 0{\cdot}0226$			$P_b^* = 0{\cdot}0058$			$P_b^* = 0{\cdot}0008$		

Consequently
$$P^*(b \mid 11, 10, 11) \leqslant 0{\cdot}05 \quad \text{when} \quad b \leqslant 2,$$
$$P^*(b \mid 11, 10, 10) \leqslant 0{\cdot}025 \quad \text{when} \quad b \leqslant 2,$$
$$P^*(b \mid 11, 10, 9) \leqslant 0{\cdot}01 \quad \text{when} \quad b \leqslant 1,$$
and
$$P^*(b \mid 11, 10, 8) \leqslant 0{\cdot}005 \quad \text{when} \quad b = 0.$$

These values of b, 2, 2, 1, and 0, are therefore tabulated in § 2 under the appropriate probabilities. The P_b^* corresponding to each tabulated b, which is usually less than the significance level, is shown in small type; for any smaller b, P_b^* will be even lower. From the existing calculations, of course, significance levels for any other selected probability less than 0·05 could easily be read, but 0·05, 0·025, 0·01 and 0·005 seem sufficient for the present.

The standard tables of the χ^2 distribution (Fisher & Yates, 1948, Table IV), provide a two-tail test when applied to a 2×2 contingency table. That is to say, on the null hypothesis that the proportion of successes in the two populations compared are equal, and assuming that frequencies are sufficiently large for the sampling distributions of the proportions to be taken as normal, the test is based on the probability of obtaining a difference as great as or greater than $(a/A - b/B)$ *in either direction* under conditions of random sampling. If a single-tail test is wanted, it can be obtained simply by entering the χ^2 table in the column corresponding to twice the level of significance. For small frequencies, however, the discrete nature of the distribution for configurations with fixed marginal frequencies cannot be ignored, and unless either $A = B$ or $2n = (A+B)$ this distribution is not symmetrical. The exact test of significance described at the beginning of this section is a single-tail test, being

based on the probability of a deviation from proportionality as great as or greater than that observed and in the direction of the observed deviation. In general, no deviation in the opposite direction will have exactly the same probability. For example, for the configuration

10	0	10
4	5	9
14	5	19

the probability is $$P^* = P = \frac{10!\,9!\,14!\,5!}{19!\,10!\,4!\,5!} = 0{\cdot}0108.$$

No deviation in the opposite direction can be regarded as the equivalent of this either in the sense of having the same difference in observed proportions or in the sense of having the same probability; the extreme configuration at the other tail is

5	5	10
9	0	9
14	5	19

for which $$P^* = P = 0{\cdot}0217.$$

If a two-tail test is wanted when the frequencies are small, a new convention must first be introduced. The only satisfactory procedure, which is consistent with the practice for larger frequencies (when the χ^2 distribution is used), is to regard the two-tail probability as given by $2P^*$. In other words, the two-tail significance tests corresponding to single-tail tests at the conventional probability levels 0·05 and 0·01 are given by comparing P^* with 0·025 and 0·005 respectively. It is this consideration which governed the choice of probability levels for tabulation in § 2.

4. Summary

This paper contains a table from which a test of significance of the deviation from proportionality in a 2 × 2 contingency table, based upon the Fisher-Yates exact probability method, can be read directly. Any 2 × 2 contingency table for which neither member of one pair of marginal totals exceeds 15 can be tested in this way, and a simple rule extends the applicability of the table to certain contingency tables with larger frequencies. The method of construction of the table is described.

I am indebted to a number of past and present members of my staff for the extensive and tedious calculations on which the table is based, especial thanks being due to Miss M. Callow.

REFERENCES

Aitken, A. C. (1944). *Statistical Mathematics*, 3rd ed. Edinburgh: Oliver and Boyd.

Barnard, G. A. (1947). Significance tests for 2 × 2 tables. *Biometrika*, **34**, 123.

Cochran, W. G. (1942). The χ^2 correction for continuity. *Iowa St. Coll. J. Sci.* **16**, 421.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Fisher, R. A. (1946). *Statistical Methods for Research Workers*, 10th ed. Edinburgh: Oliver and Boyd.

Fisher, R. A. & Yates, F. (1948). *Statistical Tables for Biological, Agricultural and Medical Research*, 3rd ed. Edinburgh: Oliver and Boyd.

Pearson, E. S. (1947). The choice of statistical tests illustrated on the interpretation of data classed in a 2 × 2 table. *Biometrika*, **34**, 139.

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *J. Roy. Statist. Soc.* Suppl. **1**, 217.

THE POWER FUNCTION OF THE TEST FOR THE DIFFERENCE BETWEEN TWO PROPORTIONS IN A 2×2 TABLE

By P. B. PATNAIK, *University College, London*

1. Introduction

Neyman & Pearson's (1933) conception of the power of a test of a statistical hypothesis, H_0, was developed, in the first instance as a means of guiding the choice between alternative tests. This, it was shown, could be done by comparing the effectiveness of the tests in discriminating between H_0 and a set of admissible alternative hypotheses regarded as most relevant to the question under test. Where there is no doubt about the most appropriate test and no sequential scheme of sampling is possible, the power function may play a useful part in indicating, before the data are collected, how large the samples should be to avoid an inconclusive result. If this procedure is to be easily applied, a ready means must be available of calculating the power of the test for a given significance level and sample size. The tables of the power function of the t-test (Neyman & Tokarska, 1936) and Tang's Tables (1938) applicable in the Analysis of Variance, are examples of such aids. The present paper aims at providing in simple, if approximate, form a means of determining the power function of the test for the difference between two proportions.

The test may be briefly outlined as follows. In two 'infinite' populations the proportions of individuals possessing a character A are $p_1(A)$ and $p_2(A)$ respectively. Random samples of m and n are drawn from the two populations and the result is represented in a 2×2 table, thus:

Table 1. *Number of individuals*

	With A	Without A	Total
Sample from first population	a	c	m
Sample from second population	b	d	n
Total	r	s	N

We may then wish to apply the test in two forms:

(i) The 'two-sided' form. To test H_0 that $p_1 = p_2$, bearing in mind the two-sided alternatives that $p_1 < p_2$ and $p_1 > p_2$.

(ii) The 'one-sided' form. To test H_0 that $p_1 \leqslant p_2$, bearing in mind the one-sided alternatives $p_1 > p_2$. These will be referred to in the following sections as cases (i) and (ii).

In obtaining a solution we shall regard the sample space as two-dimensioned as in Pearson's (1947) Problem II or Barnard's (1947) '2×2 comparative trial'. Fig. 1 (*a*) roughly illustrates the sample space for the case $m = n = 50$. A possible sampling result is represented by a point (a, b) in this lattice of $51 \times 51 = 2601$ points. Sample points having $a + b = r$ are

said to lie on the 'diagonal' r of the lattice. In the case of small samples, or even for large samples when (a, b) falls near the $(0, 0)$ or (m, n) corners of the lattice, considerable difficulties arise owing to the discontinuity of the distribution. We shall be concerned in the first place with examining the case where it is justifiable to apply the test by referring to the normal probability scale the ratio

$$\left.\begin{aligned} u &= \frac{a - \frac{rm}{N}}{\sqrt{\frac{mnrs}{N^2(N-1)}}} \\ &= \frac{\frac{a}{m} - \frac{b}{n}}{\sqrt{\left[\frac{r}{N}\left(1-\frac{r}{N}\right)\left(\frac{1}{m}+\frac{1}{n}\right)\right]}} = \frac{(ad-bc)\sqrt{N}}{\sqrt{(mnrs)}} \quad \text{(approx.)}, \end{aligned}\right\} \tag{1}$$

where the second and third alternative forms are equivalent to the first if the factor $N-1$ is replaced by N. The first form brings out the fact that u is the ratio of the deviation from the mean to the standard deviation, in the hypergeometric series representing the conditional distribution of a for r fixed. The second form arises from the classical approach in which a difference of observed proportions is compared with an estimate of its standard error. The third, when squared, gives the commonly quoted form of the expression for χ^2, with one degree of freedom, in a 2×2 table.

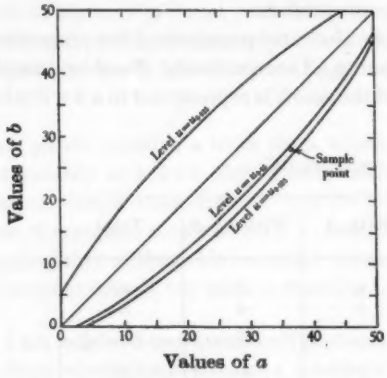


Fig. 1 (*a*). Sample space ($m = n = 50$).

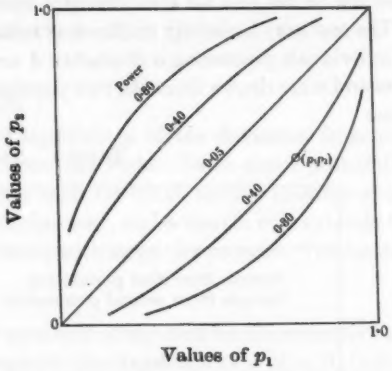


Fig. 1 (*b*). Power contours ($m = n = 50$).

Write u_α for the 100α percentage point of the normal distribution $N(0, 1)$ i.e.

$$\int_{u_\alpha}^{\infty} \frac{1}{\sqrt{(2\pi)}} e^{-\frac{1}{2}u^2} = \alpha,$$

then using the two-sided test (i), we should reject the hypothesis that $p_1 = p_2$ at the significance level α when $|u| > u_{\frac{1}{2}\alpha}$; in the case of the one-sided test, we should reject the hypothesis that $p_1 \leqslant p_2$ at the same significance level when $u > u_\alpha$. Thus, in the example illustrated in Fig. 1 (*a*), it is seen from the position of the sample point, that H_0 would not be rejected at the 5 % level if the test is in form (i), but would be rejected at the 5 % level if it is in form (ii).

The boundaries of the critical region associated with the test, taking say, case (i), are formed by obtaining from the equation

$$\frac{a-\frac{mr}{N}}{\sqrt{\frac{mnrs}{N^2(N-1)}}} = u_{\frac{1}{2}\alpha},$$

the 'cut-off' points $(a, r-a)$ on each diagonal r in the lattice and joining them as in Fig. 1 (a). Subject to the error involved in the approximation, the chance is α that the sample point falls in the critical region when p_1 and p_2 have a common, though unknown, value i.e. in Neyman & Pearson's notation, $P\{E \in w_\alpha \mid p_1 = p_2\} = \alpha$. If in fact $p_1 \neq p_2$, then the power of the test of H_0 with regard to the alternative hypothesis $H_1(p_1, p_2)$ is the chance that the point (a, b) falls in the critical region when sampling from populations with proportions p_1 and p_2. Or, formally,

$$P\{E \in w_\alpha \mid p_1, p_2\} = P\{|u| > u_{\frac{1}{2}\alpha} \mid p_1, p_2\} \quad \text{for case (i),}$$

or

$$= P\{u > u_\alpha \mid p_1, p_2\} \quad \text{for case (ii).}$$

This is the total probability density at all the discrete points (a, b) included in the critical region. The problem is to express this in a readily calculable form.

If this is done, two types of application are evident:

(1) When the decision has been made to take two samples of, say, 50, or when the available data happen to consist in samples of this size, we may ask, 'what is the chance that the test described will show a difference in observed proportions significant at the 5 % level when in fact, p_1 and p_2 are as different, say, as 0·50 and 0·65?'

(2) On the other hand, we may use the theory to ask in advance how large the samples should be so that the risk of failing to detect a given difference* between p_1 and p_2 which is considered to be of importance, shall be acceptably small. For example, we may ask, 'what sizes of samples should we take so that in applying our test we may have a high, say a 90 %, chance of detecting that the proportions are not equal when, in fact, they are as different as 0·50 and 0·65?'

It is clear that for given m, n and α the power of the test will be constant on certain contours in the p_1, p_2 space such as those shown in Fig. 1 (b). We shall examine the approximate form of these contours and show that in the important case when $m = n = \frac{1}{2}N$, the family of contours is independent of N and α, although the power associated with a particular contour will be a function of N and α, which has been tabled. Throughout the investigation, approximations are made of the type involved in representing binomial or hypergeometric distributions by normal distributions. The adequacy of these approximations is examined in certain cases.

2. Exact values for the power function of the test based on the ratio u

Since a and b are independent, it follows that for the population proportions, p_1 and p_2

$$p(a, b) = \frac{m!}{a!\,c!} p_1^a q_1^c \times \frac{n!}{b!\,d!} p_2^b q_2^d \quad (q_1 = 1-p_1,\ q_2 = 1-p_2). \tag{2}$$

If our test consists in rejecting $H_0(p_1 = p_2)$ when $|u| > u_{\frac{1}{2}\alpha}$, its power with respect to an alternative $H_1(p_1 \neq p_2)$ is the sum of the values of $p(a, b)$ for all points (a, b) lying in the critical region. Exact calculation is laborious, since it involves the multiplication of terms of one binomial by the sums of terms of the other, which may be obtained from the tables of the

* Neyman & Pearson have termed this the error of the second kind.

Incomplete Beta Function. In the special case $m = 18, n = 12$, discussed in Pearson's paper, exact values of the power function have been calculated for $p_1, p_2 = 0{\cdot}1(0{\cdot}1)0{\cdot}9$ for the two-sided test and for significance levels 10 and 2 %; they are tabulated in Tables 2(*a*) and

Table 2(*a*). *Power function for case* $m = 18, n = 12$. *Significance level* $\alpha = 0{\cdot}10$

Approximate values are shown in parentheses

p_2 \ p_1	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9
0·9	1·000 (1·00)	0·997 (0·999)	0·975 (0·992)	0·907 (0·96)	0·773 (0·87)	0·574 (0·70)	0·356 (0·44)	0·170 (0·21)	0·088 (0·10)
0·8	0·997 (0·999)	0·972 (0·980)	0·882 (0·91)	0·732 (0·77)	0·494 (0·56)	0·289 (0·34)	0·144 (0·17)	0·094 (0·10)	0·202
0·7	0·977 (0·987)	0·880 (0·90)	0·722 (0·73)	0·487 (0·51)	0·272 (0·30)	0·136 (0·15)	0·092 (0·10)	0·159	0·393
0·6	0·918 (0·94)	0·742 (0·75)	0·515 (0·51)	0·287 (0·29)	0·142 (0·15)	0·099 (0·10)	0·155	0·320	0·610
0·5	0·796 (0·83)	0·534 (0·54)	0·309 (0·30)	0·156 (0·15)	0·100 (0·10)	0·156	0·309	0·534	0·796
0·4	0·610 (0·64)	0·320 (0·32)	0·155 (0·15)	0·099 (0·10)	0·142	0·287	0·515	0·742	0·918
0·3	0·393 (0·40)	0·159 (0·16)	0·092 (0·10)	0·136	0·272	0·487	0·722	0·885	0·977
0·2	0·202 (0·20)	0·094 (0·10)	0·144	0·289	0·494	0·732	0·882	0·972	0·997
0·1	0·088 (0·10)	0·170	0·356	0·574	0·773	0·907	0·975	0·997	1·000

Table 2(*b*). *Power function for case* $m = 18$, $n = 12$. *Significance level* $\alpha = 0{\cdot}02$

Approximate values are shown in parentheses

p_2 \ p_1	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9
0·9	0·998 (1·00)	0·976 (0·993)	0·902 (0·96)	0·752 (0·86)	0·532 (0·68)	0·290 (0·43)	0·103 (0·20)	0·017 (0·06)	0·006 (0·02)
0·8	0·961 (0·990)	0·882 (0·91)	0·694 (0·75)	0·490 (0·52)	0·265 (0·30)	0·115 (0·12)	0·035 (0·05)	0·016 (0·02)	0·055
0·7	0·909 (0·94)	0·714 (0·73)	0·460 (0·48)	0·248 (0·26)	0·111 (0·11)	0·041 (0·04)	0·021 (0·02)	0·050	0·177
0·6	0·767 (0·81)	0·500 (0·49)	0·262 (0·25)	0·111 (0·11)	0·041 (0·04)	0·023 (0·02)	0·048	0·138	0·361
0·5	0·625 (0·66)	0·293 (0·28)	0·124 (0·11)	0·045 (0·04)	0·022 (0·02)	0·045	0·124	0·293	0·625
0·4	0·361 (0·37)	0·138 (0·13)	0·048 (0·04)	0·023 (0·02)	0·041	0·111	0·262	0·500	0·767
0·3	0·177 (0·18)	0·050 (0·05)	0·021 (0·02)	0·041	0·111	0·248	0·460	0·714	0·909
0·2	0·055 (0·06)	0·016 (0·02)	0·035	0·115	0·265	0·490	0·694	0·882	0·961
0·1	0·006 (0·02)	0·017	0·103	0·290	0·532	0·752	0·902	0·976	0·998

2(*b*).* When $p_1 = p_2$ the value of the power function should reduce to α. But in these tables we see that the values on the diagonal, $p_1 = p_2$ are not exactly equal to 0·10 or 0·02. The discrepancy is due to the fact that continuous approximations have been made to the discontinuous distributions in the formulation of the test.

Interpolation between the calculated values in Table 2(*a*), leads to the power contours shown in Fig. 2. The test has the same power of establishing significance when sampling from any populations for which p_1, p_2 lies on a given contour. The chances of establishing significance are written alongside the contours. The surface for which the ordinates at p_1, p_2 are equal to the power, may be called the power surface of the test.

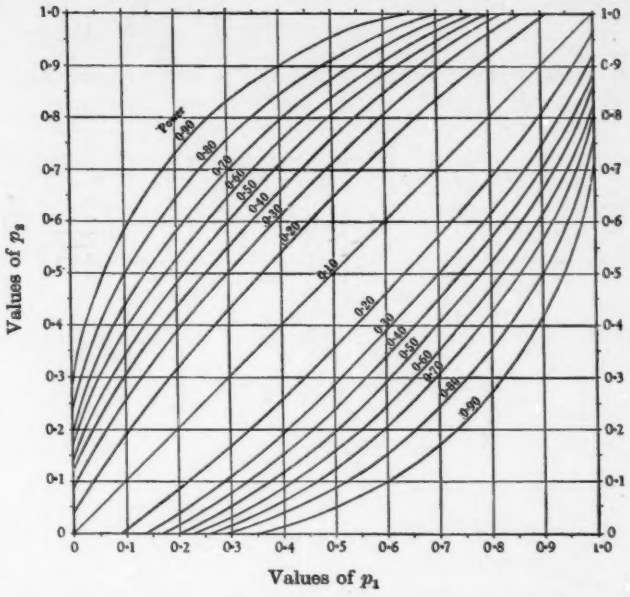


Fig. 2. Power contours (special case $m = 18$, $n = 12$) for significance level 0·10.

We can see from the contour diagram that with samples of this size (18 and 12), even when the p's are as different as 0·6 and 0·3 there is only a 50 % chance of establishing significance. If instead of the high level $\alpha = 0{\cdot}10$ employed here, a lower level were chosen, then the p's must differ even more to give the same chance of establishing significance. Thus the diagram illustrates well how inadequate small samples are to establish what would ordinarily be regarded as a difference of some importance.

It can also be seen that $|p_1 - p_2|$ is nearly constant on a power contour near the middle of the square; e.g. the difference is roughly 0·3 on the 0·5 contour within the range (0·45, 0·15) and (0·75, 0·45).

* For these calculations, I am indebted to the members of the Statistics Department of University College, London, in particular to Mr V. D. Gangolli.

3. Approximations to the power function

We will now consider an approximation to the distribution of a under the hypothesis H_1, under which the population proportions are p_1 and p_2. From (1)

$$p(a,r) = \frac{N!}{r!s!} p_2^r q_2^s \left(\frac{q_1}{q_2}\right)^m \times \left\{\frac{m!n!r!s!}{N!a!b!c!d!}\right\} \left(\frac{p_1 q_2}{p_2 q_1}\right)^a.$$

If we replace the hypergeometric term in the curled brackets by the ordinate of a normal curve having the mean and S.D. of the hypergeometric series, then

$$p(a,r) = \frac{N!}{r!s!} p_2^r q_2^s \left(\frac{q_1}{q_2}\right)^m \times \frac{1}{\sqrt{(2\pi)}\sqrt{\dfrac{mnrs}{N^2(N-1)}}} \exp\left[-\frac{\left(a-\dfrac{rm}{N}\right)^2}{2\dfrac{mnrs}{N^2(N-1)}}\right] \times \left(\frac{p_1 q_2}{p_2 q_1}\right)^a.$$

Writing $\left(\dfrac{p_1 q_2}{p_2 q_1}\right)^a$ as $\exp\left[a \log_e \left(\dfrac{p_1 q_2}{p_2 q_1}\right)\right]$, collecting the terms containing a and making a perfect square, we obtain $p(a,r) = p(r) \times p(a \mid r)$,

where
$$p(r) = \frac{N!}{r!s!} p_2^r q_2^s \left(\frac{q_1}{q_2}\right)^m \exp\left[\frac{rm}{N} \log_e \frac{p_1 q_2}{p_2 q_1} + \frac{mnrs}{2N^2(N-1)} \left(\log_e \frac{p_1 q_2}{p_2 q_1}\right)^2\right], \tag{3}$$

and
$$p(a|r) = \frac{1}{\sqrt{(2\pi)}\sqrt{\dfrac{mnrs}{N^2(N-1)}}} \exp\left[-\frac{\left(a-\dfrac{rm}{N}-\dfrac{mnrs}{N^2(N-1)} \log_e \dfrac{p_1 q_2}{p_2 q_1}\right)^2}{2\dfrac{mnrs}{N^2(N-1)}}\right]. \tag{4}$$

Thus (4) is the approximate conditional distribution of a on the diagonal $r = a+b$ and is seen to be normal, with

$$\text{mean} = \frac{rm}{N} + \frac{mnrs}{N^2(N-1)} \log_e \frac{p_1 q_2}{p_2 q_1}$$

and
$$\text{S.D.} = \sqrt{\frac{mnrs}{N^2(N-1)}}.$$

Defining
$$u = \frac{a - \dfrac{rm}{N}}{\sqrt{\dfrac{mnrs}{N^2(N-1)}}}$$

as in (1), equation (4) becomes

$$p(u|r) = \frac{1}{\sqrt{(2\pi)}} \exp\left[-\frac{1}{2}\left(u - \sqrt{\frac{mnrs}{N^2(N-1)}} \log_e \frac{p_1 q_2}{p_2 q_1}\right)^2\right]$$

$$= \frac{1}{\sqrt{(2\pi)}} e^{-\frac{1}{2}(u-h(r))^2}, \tag{5}$$

where
$$h(r) = \sqrt{\frac{mnrs}{N^2(N-1)}} \log_e \frac{p_1 q_2}{p_2 q_1}, \tag{6}$$

and is a function of r only, since $s(=N-r)$ and the other quantities are given.

If p_1 and p_2 are equal, then (5) reduces to the distribution of u under H_0,

$$p(u) = \frac{1}{\sqrt{(2\pi)}} e^{-\frac{1}{2}u^2}, \tag{7}$$

which is the normal approximation used in obtaining the test criterion. This distribution is independent of r; it is this fact that makes the critical regions on each diagonal r similar regions which can be combined into a single region in the two-dimensioned sample space.

But the distribution of u under H_1, given by (5) is not independent of r. It is normal, with the same S.D. as for (7), i.e. unity, but with its mean shifted by $h(r)$.

What may be termed the 'conditional power', for r fixed, with regard to $H_1(p_1 \neq p_2)$ is then for case (i),

$$P(|u| > u_{\frac{1}{2}\alpha}|r) = \int_{-\infty}^{-u_{\frac{1}{2}\alpha}} p(u|r)\,du + \int_{u_{\frac{1}{2}\alpha}}^{\infty} p(u|r)\,du = 1 - \frac{1}{\sqrt{(2\pi)}} \int_{-u_{\frac{1}{2}\alpha}-h(r)}^{u_{\frac{1}{2}\alpha}-h(r)} e^{-\frac{1}{2}u^2} du, \tag{8}$$

and for case (ii), $$P(u > u_\alpha|r) = \int_{u_\alpha}^{\infty} p(u|r)\,du = \frac{1}{\sqrt{(2\pi)}} \int_{u_\alpha - h(r)}^{\infty} e^{-\frac{1}{2}u^2} du. \tag{9}$$

Since $p(u) = p(u|r)\,p(r)$, the 'over-all' power function or, simply, the power function is

$$\int_{-\infty}^{\infty} P(|u| > u_{\frac{1}{2}\alpha}|r)\,p(r)\,dr \tag{10}$$

for case (i) with a similar expression for case (ii). It is clear that even with the simplified expression (8) for the conditional power function, depending on $h(r)$, the labour involved in calculating the over-all power (10) would in general be prohibitive. Some approximation is therefore required for $p(r)$, i.e. the expression in (3). The simplest approximation is obtained by assuming r to be normally distributed. Since a and b are distributed binomially with means mp_1, and np_2 and S.D.'s mp_1q_1 and np_2q_2, $r = a+b$ can be considered as distributed normally with mean $= mp_1 + np_2$ and S.D. $= \sqrt{(mp_1q_1 + np_2q_2)}$. That is,

$$p(r) = \frac{1}{\sqrt{(2\pi)}\sqrt{(mp_1q_1 + np_2q_2)}} \exp\left[-\frac{(r-(mp_1+np_2))^2}{2(mp_1q_1+np_2q_2)}\right]. \tag{11}$$

Hence, the expression (10) for the over-all power function becomes

$$\frac{1}{\sqrt{(2\pi)}\sqrt{(mp_1q_1+np_2q_2)}} \int_{-\infty}^{\infty} P(|u| > u_{\frac{1}{2}\alpha}|r) \times \exp\left[-\frac{1}{2}\frac{(r-(mp_1+np_2))^2}{mp_1q_1+np_2q_2}\right] dr \tag{12}$$

for case (i) with a similar expression for case (ii).

Though problems may arise where the conditional power function would be useful, clearly it is the value of the over-all power that will help in determining in advance of the experimental result how large the difference between p_1 and p_2 must be for the standard test to have a given chance of establishing significance. For in such a preliminary survey, r cannot be regarded as fixed.

4. EVALUATION OF OVER-ALL POWER

The over-all power is a function of p_1 and p_2, for given m, n and α and may be written as $\beta(p_1, p_2 | m, n, \alpha)$. Similarly, the conditional power function may be written as

$$\beta(p_1, p_2 | m, n, \alpha, r).$$

They will be denoted here for simplicity by β and $\beta(r)$ respectively.

Suppose $\mu_1 = mp_1 + np_2$, $\sigma^2 = \mu_2 = mp_1q_1 + np_2q_2$ and μ_3, μ_4, etc. equal the higher moments of the normal distribution (11). Then for case (i) or case (ii), from (12)

$$\beta = \frac{1}{\sqrt{(2\pi)}\,\sigma} \int_{-\infty}^{\infty} \exp\left[-\frac{(r-\mu_1)^2}{2\sigma^2}\right] \beta(r)\,dr.$$

Expanding $\beta(r)$ by Taylor's Theorem,

$$\beta(r) = \beta(\mu_1) + (r-\mu_1)\beta'(\mu_1) + \frac{(r-\mu_1)^2}{2!}\beta''(\mu_1) + \ldots \tag{13}$$

and substituting in above, we obtain

$$\beta = \beta(\mu_1)\frac{1}{\sqrt{(2\pi)}\,\sigma}\int_{-\infty}^{\infty}\exp\left[-\frac{(r-\mu_1)^2}{2\sigma^2}\right]dr$$
$$+\beta'(\mu_1)\frac{1}{\sqrt{(2\pi)}\,\sigma}\int_{-\infty}^{\infty}\exp\left[-\frac{(r-\mu_1)^2}{2\sigma^2}\right](r-\mu_1)\,dr$$
$$+\frac{\beta''(\mu_1)}{2!}\frac{1}{\sqrt{(2\pi)}\,\sigma}\int_{-\infty}^{\infty}\exp\left[-\frac{(r-\mu_1)^2}{2\sigma^2}\right](r-\mu_1)^2\,dr+\ldots$$
$$=\beta(\mu_1)+\frac{\beta''(\mu_1)}{2!}\mu_2+\frac{\beta^{(4)}(\mu_1)}{4!}\mu_4+\ldots. \tag{14}$$

It follows that a first approximation to the over-all power β is

$$\beta(\mu_1) = 1-\frac{1}{\sqrt{(2\pi)}}\int_{-u_{\frac{1}{2}\alpha}-h(\mu_1)}^{u_{\frac{1}{2}\alpha}-h(\mu_1)} e^{-\frac{1}{2}u^2}\,du \quad \text{for case (i)} \tag{15}$$

or

$$=\frac{1}{\sqrt{(2\pi)}}\int_{u_\alpha-h(\mu_1)}^{\infty} e^{-\frac{1}{2}u^2}\,du \quad \text{for case (ii)} \tag{16}$$

substituting μ_1 for r in the expressions (8) and (9).

A second approximation will be derived and discussed later in section 10.

5. COMPARISON OF EXACT AND APPROXIMATE VALUES OF THE POWER FUNCTION

Taking the two-sided test, its over-all power has been considered in three special cases. Using the first approximation, the values of β have been calculated for $m = 18, n = 12$ and are shown in parentheses below the exact values in Tables 2(*a*) and 2(*b*). The exact and the approximate values have also been calculated for $m = n = 15$ and $m = n = 30$ for a few combinations of p_1, p_2, selected so as to give high power which, as we shall see later, is in the range we are most interested in. These are given in Tables 3(*a*) and 3(*b*). In the latter case, since N is large, $N-1$ is replaced by N in the ratio u.

Table 3(*a*). *Showing the power of the two-sided test.* $m = n = 15$

p_1	p_2	Significance level, α					
		0·10			0·02		
		Exact value	First approx.	Second approx.	Exact value	First approx.	Second approx.
0·3	0·4	0·141	0·16	0·154	0·034	0·04	0·041
0·6	0·8	0·306	0·34	0·334	0·112	0·14	0·133
0·1	0·3	0·389	0·44	0·422	0·149	0·20	0·193
0·2	0·7	0·896	0·92	0·912	0·680	0·76	0·750
0·05	0·5	0·916	0·97	0·964	0·736	0·90	0·876
0·1	0·6	0·919	0·96	0·953	0·739	0·86	0·844
0·2	0·8	0·974	0·98	0·982	0·872	0·93	0·923
0·1	0·7	0·980	0·992	0·991	0·894	0·96	0·954

Table 3(*b*). *Showing the power of the two-sided test.* $m = n = 30$

p_1	p_2	Significance level, α					
		0·10			0·02		
		Exact value	First approx.	Second approx.	Exact value	First approx.	Second approx.
0·05	0·3	0·884	0·93	0·904	0·631	0·78	0·752
0·1	0·4	0·885	0·91	0·903	0·691	0·75	0·736
0·3	0·7	0·937	0·95	0·947	0·807	0·83	0·824
0·2	0·6	0·945	0·96	0·957	0·839	0·86	0·852
0·1	0·5	0·977	0·988	0·985	0·902	0·94	0·934
0·2	0·7	0·993	0·996	0·996	0·965	0·98	0·974

From these tables it can be seen generally that (1) the first approximation over-estimates the power, (2) the agreement is better with large sample sizes, (3) the discrepancy is less when p_1 and p_2 are near 0·5 than when one or both is very small, i.e. when we are in a corner of the p_1, p_2-space, (4) the approximation is better for $\alpha = 0{\cdot}10$ than for $\alpha = 0{\cdot}02$.

Mathematically, the approximation is of course not very accurate even with two samples of 30. But it must be remembered (*a*) that what is needed in practice is a simple procedure for determining quickly the chance of establishing significance, and (*b*) that errors of approximation are not everywhere of the same importance. When the chance of detecting a worthwhile difference between p_1 and p_2 is only 0·40 it is clear that the sample size is inadequate and this would still be our conclusion if the approximation gave 0·45. It is perhaps only when the power approaches 0·90 that an error of this order becomes serious. If the true chance were 0·90 (odds of 9 to 1) and the approximation gave 0·95 (odds of 19 to 1) the result becomes somewhat misleading on this basis. Examination of Tables 3(*a*) and 2(*b*) suggests that if one value of p is likely to be less than 0·1 or greater than 0·9, the approximation is failing us for $n < 30$. But even here, as shown on p. 170 below when estimating the sample size needed to provide a given power, using Tables 4 and 5, we shall not be far out.

6. Case of equal sample sizes

Suppose $m = n = \frac{1}{2}N$. Then $\mu_1 = n(p_1 + p_2)$. Substituting in (6) and replacing $N^2(N-1)$ by $N^3 (= 8n^3)$, the error being negligible when n is not too small, we find

$$h(\mu_1) = \sqrt{n}\frac{\sqrt{2}}{4}\sqrt{[(p_1+p_2)(2-p_1-p_2)]}\log_e\frac{p_1(1-p_2)}{p_2(1-p_1)}.$$

In the case of the one-sided test, where the alternative is $p_1 > p_2$, $h(\mu_1)$ is positive. In the other case with alternatives $p_1 < p_2$ or $p_1 > p_2$, $h(\mu_1)$ is negative or positive. But from the expression (15) for the approximate power, it is seen that the sign of $h(\mu_1)$ is immaterial. So, putting $h = |h(\mu_1)|$,

$$k = k(p_1, p_2) = \frac{\sqrt{2}}{4}\sqrt{[(p_1+p_2)(2-p_1-p_2)]}\left|\log_e\frac{p_1(1-p_2)}{p_2(1-p_1)}\right|, \tag{17}$$

we have

$$h = k\sqrt{n}. \tag{18}$$

From (15) and (16) it follows that to the first approximation, for given $n = \frac{1}{2}N$ and α, the power, β, is a function of k only, i.e.

$$\beta(p_1, p_2 \,|\, n, n, \alpha) = \beta(k \,|\, n, \alpha).$$

These contours of constant k have been drawn in Fig. 3 and the values of k are written alongside. Thus for $m = n$ the values of p_1, p_2 and the power of the test may be linked up as follows:

(i) Fig. 3 relates p_1, p_2 to k.

(ii) Equation (18) gives h in terms of k and n.

(iii) The normal integrals (15) and (16) give the power in terms of h and the significance level α employed in the test.

It will be seen that to the first approximation the composite hypothesis,

$$H_0(p_1 = p_2 = p, \text{ unknown}),$$

is reduced to a simple hypothesis, $k = 0$. So the test in this case may be regarded as the test of the hypothesis, $k = 0$, with alternatives, $k > 0$.

7. Tabulation

(1) *Tables of k as an alternative to Fig.* 3. From (17), k has been calculated for

$$p_1, p_2 = 0{\cdot}05(0{\cdot}05)\,0{\cdot}95$$

and is given in Table 4 (printed at the end of the paper, p. 174). Since $k(1-p_1, 1-p_2)$ has the same value as $k(p_1, p_2)$ the figures in the upper part of the table ($p_2 > p_1$) are not printed.

(2) *Table* 5. *The power as a function of* $h = k\sqrt{n}$ *and* α. For the two-sided test we require the integral (15). Denoting this by P, i.e.,

$$P = 1 - \frac{1}{\sqrt{(2\pi)}} \int_{-u_{\frac{1}{2}\alpha}-h}^{u_{\frac{1}{2}\alpha}-h} e^{-\frac{1}{2}u^2}\, du, \tag{19}$$

the value of P is given in columns 2–5 of Table 5 (printed at the end of the paper, p. 175). It has been calculated from *Tables of Probability Functions*, Vol. II (Federal Works Agency, New York), using Lagrangian four-point interpolation, for the levels

$$\alpha = 0{\cdot}10, 0{\cdot}05, 0{\cdot}02 \text{ and } 0{\cdot}01 \quad \text{and for} \quad h = 0{\cdot}1(0{\cdot}1)\,3{\cdot}0(0{\cdot}2)\,5{\cdot}0.$$

As h increases, the contribution to this integral from one tail rapidly becomes negligible, so that (19) approximates to

$$\frac{1}{\sqrt{(2\pi)}} \int_{u_{\frac{1}{2}\alpha}-h}^{\infty} e^{-\frac{1}{2}u^2}\, du.$$

From (16), we see that this integral is the power of the one-sided test, applied at the significance level $\frac{1}{2}\alpha$. The points at which this result is obtained to a given level of accuracy are shown in Table 5; at least two-decimal accuracy occurs below the mark *, three-decimal accuracy below †, and four below ‡.

To facilitate the use, values of $k = h/\sqrt{n}$ have been tabulated in the right-hand side of Table 5 corresponding to the h values in the first column and for

$$n = 10(5)\,50(10)\,100, 150.$$

Linear interpolation is adequate for Tables 4 and 5, except in the corners of Table 4 where Lagrangian four-point interpolation may be necessary.

8. APPLICATION OF THE POWER FUNCTION AND USE OF TABLES

Illustration 1. Laurence & Newell, while experimenting with the composition of soil composts, recorded the following results of a germination trial with *Primula sinensis* seeds (quoted in *Statistical Analysis in Biology* by K. Mather, 1946, p. 193). Two equal groups of seeds were allowed to germinate in dishes containing filter papers soaked respectively in rain water and in water allowed to seep through loam before use.

Table 6

	Germinated	Ungerminated	Total
Loam water	37	13	50
Rain water	32	18	50
Total	69	31	100

To test if the type of water affects germination, u has been calculated to be 1·11 and referred to the normal probability scale. It is seen that there is no significant difference at the 5 % level.

It might be asked what magnitude of difference could we hope to detect using two samples of 50. Suppose, for example, that for these populations 80 %, say, of seeds will germinate in loam water and 60 % in rain water; what would have been the chance of establishing significance at 5 % level?

To obtain this, we find from Fig. 3 or from Table 4 the value of k for $p_1 = 0{\cdot}8$ and $p_2 = 0{\cdot}6$. From Table 4 it is seen to be 0·318. Then we enter Table 5 in the column of $n = 50$ and find that this value of k lies between the tabulated values, 0·311 and 0·325. They correspond to the figures 0·5949 and 0·6331 in the column of P for $\alpha = 0{\cdot}05$. So, the first approximation to the power lies between these values and by linear interpolation is found to be 0·61. If the level chosen for the test is $\alpha = 0{\cdot}01$, we find that the power is only 0·37. Clearly this indicates that with two samples of 50, there is a very considerable risk of failing to establish significance when the difference in chances of germination is of this order.

Suppose now we asked how large the samples should have been to give a chance of 0·9 of establishing significance when the true percentages germinating are 80 in loam water and 60 in rain water? We then proceed as follows: In Table 5, entering the column of P under $\alpha = 0{\cdot}05$, we find that 0·90 lies between 0·8925 and 0·9251 and following the rows of these figures we see that $k = 0{\cdot}318$ lies between the figures in the columns of $n = 100$ and $n = 150$. As the interval is too wide for interpolation we find h in column 1 corresponding to $P = 0{\cdot}90$ in column 3 and then from the relation, $n = h^2/k^2$, we obtain n to be nearly 105. If the level chosen is $\alpha = 0{\cdot}01$, the samples should be of size 150 to give the same power.

Illustration 2. In the early stages of production of an important piece of electrical equipment, it has been found that the percentage of units failing under test varies, according to the batch, between 30 and 40 %. An adjustment is suggested which would be considered worthwhile if it leads to a 50 % reduction in failures. A trial is planned in which $N = 2n$ units are selected at random from a batch, half are adjusted and half not and the whole are then tested. We want to know in advance about how large should n be so that the odds are

19 to 1 that we shall not reach an inconclusive result (at the 5 % level) if there has been a 50 % reduction in failures.

Here $p_1 = 0{\cdot}30$ to $0{\cdot}40$ and $p_2 = 0{\cdot}15$ to $0{\cdot}20$. From Fig. 3 we see that the contour, $k = 0{\cdot}3$ roughly passes through the range of these points (p_1, p_2). From Table 5 we find $h = 3{\cdot}3$ for $P = 0{\cdot}95$ and $\alpha = 0{\cdot}10$ (since the test is of the one-sided form). So, $n = h^2/k^2 = 120$, nearly, a surprisingly large number.

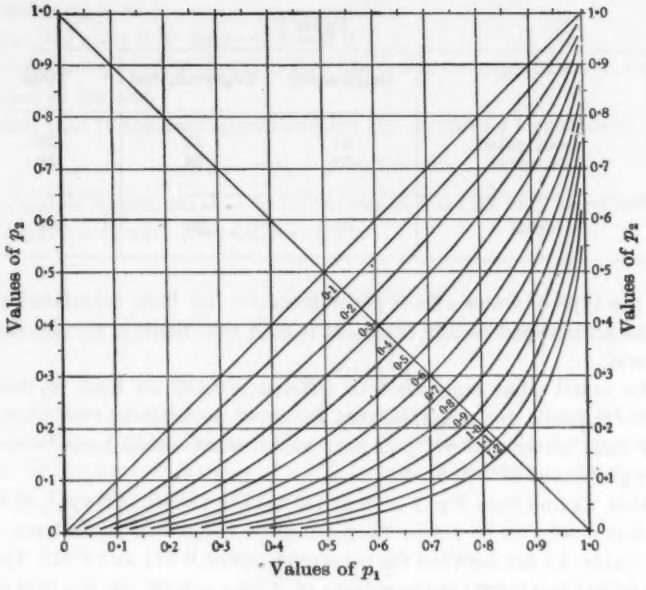


Fig. 3. Contours of constant k for use in determining the power when $m = n$. (Values of k are printed alongside the curves).

Sometimes, without precisely specifying p_1 or p_2 we may consider an improvement worthwhile if it increases the percentage of effectives by a fixed amount, e.g. such that $p_2 - p_1 = 0{\cdot}25$. Table 4 shows how this difference is roughly constant for k constant in the central area; for example,

p_1	p_2	k
0·20	0·45	0·393
0·25	0·50	0·376
0·35	0·60	0·362
0·45	0·70	0·366
0·55	0·80	0·393

That is, the contour $k = 0{\cdot}38$ passes very closely through these points (p_1, p_2). With this value of k we obtain n as before.

Illustration 3. If we have a random variable x following a distribution law which is approximately normal, the most efficient estimator of the population mean, μ, is the sample

mean $\bar{x}$. On the assumption that the variance is not changing appreciably, we should use the t-test to determine whether there has been a change in μ between the drawing of a first and second sample. Practical requirements sometimes make it preferable, or even necessary, to observe only the number of individuals in the two samples, say a and b respectively, for which x falls below a fixed level x_0. We can then test whether there has been a change in μ with the help of the ratio u of equation (1), p_1 and p_2 being the chance that $x \leqslant x_0$ for the first and second samples, respectively.

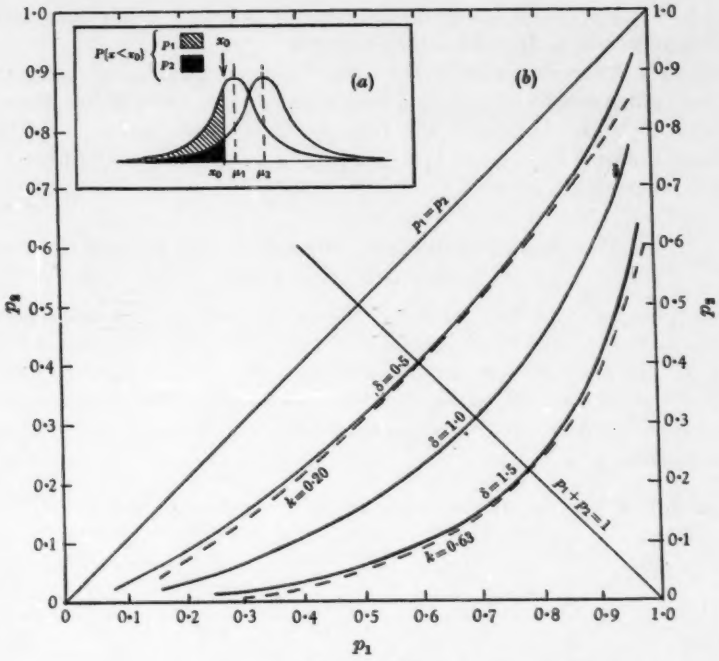


Fig. 4 (*a*), (*b*).

Examples of this problem occur:

(i) In firing two types of shot, with striking velocity x_0, against a standard proof plate and observing the numbers, a and b, that fail to perforate.

(ii) In dosage-mortality problems, where two drugs are compared only at a single dosage level, x_0.

The value of x_0 will frequently be at our choice and the generally accepted principle is to take a value so that $P\{x \leqslant x_0\}$ is in the neighbourhood of 0·5. It is possible to confirm the soundness of this procedure in terms of the power function of the u (or χ^2) test. Fig. 4 (*a*) represents two normal distributions with means μ_1 and μ_2 and a common standard deviation, σ. p_1 and p_2 are the proportionate areas under the curves below $x = x_0$. For a given relative shift in mean, $\delta = (\mu_2 - \mu_1)/\sigma$, values of p_1 and p_2 as functions of x_0 are readily found from tables of the normal probability integral. Fig. 4 (*b*) shows, as solid line curves, the locus of points p_1, p_2 for which $\delta = 0{\cdot}5$, 1·0 and 1·5. Our problem is to determine, for given δ, the

value of x_0 which will maximize the chance that the u-test will establish significance. We know that in large samples of equal size ($m = n$), the power of the test is constant on the contours of constant k (equation (17)). Two of these contours, for $k = 0{\cdot}20$ and $0{\cdot}63$, approximately, are shown as dotted lines in Fig. 4(*b*); they touch the curves for $\delta = 0{\cdot}5$ and $1{\cdot}5$, respectively, at the points where $p_1 + p_2 = 1$, and elsewhere fall beyond them, as shown.

Since for small samples, we have found that the k-contours rather overestimate the power when p_1 or p_2 approach 0 or 1, it follows that the true power contour which touches a δ-curve where $p_1 + p_2 = 1$, will fall even further outside it, towards the corners of the diagram, than the k-contours, as drawn. It appears therefore that:

(i) For a given δ, the usual test for a difference in proportions has a maximum chance of establishing a difference if $x_0 = \frac{1}{2}(\mu_1 + \mu_2)$ and this is so for all levels of the power function.*

(ii) When $m = n$ are large and the true power contours become exactly those of $k =$ constant, shown in Figs. 3 and 4(*b*), $p_1 + p_2$ may differ considerably from 1·0 without appreciably reducing the power of the test.

9. The error in estimating the sample size necessary to ensure a given power

In section 5 it has been seen that the power obtained by the first approximation is generally an over-estimate of the true value. The effect of this will be to underestimate n in carrying out the procedure illustrated in the previous example. It is possible to determine the magnitude of this error in the neighbourhood of $n = 15$ and 30, where the exact values of the power have been found and are given in Tables 3(*a*) and (*b*). Some results of this comparison are shown in Table 7.

Table 7. *Comparing the estimates with the true values of the sample size, n*

$m = n$	p_1	p_2	True power*		k from Table 4	h for true power from Table 5		Estimate of $n = h^2/k^2$	
			For $\alpha = 0{\cdot}10$	For $\alpha = 0{\cdot}02$		$\alpha = 0{\cdot}10$	$\alpha = 0{\cdot}02$	$\alpha = 0{\cdot}10$	$\alpha = 0{\cdot}02$
15	0·05	0·5	0·916	0·736	0·930	3·03	2·96	11	11
	0·1	0·6	0·919	0·739	0·878	3·05	2·97	13	12
	0·1	0·7	0·980	0·894	1·055	3·71	3·58	13	12
30	0·05	0·3	0·884	0·631	0·563	2·84	2·66	26	23
	0·1	0·5	0·977	0·902	0·712	3·65	3·62	27	26
	0·2	0·7	0·993	0·965	0·786	4·12	4·15	28	28

* The power of the two-sided test.

For example, if $p_1 = 0{\cdot}1$, $p_2 = 0{\cdot}6$, Table 3(*a*) gives the true power for $m = n = 15$ as (i) 0·919 using the test with a 10 % significance level and (ii) 0·739 using the test with a 2 % level. Suppose now that we were to use Tables 4 and 5 to estimate how large the sample

* For $m = n = 30$, Table 5 shows that for $k = 0{\cdot}20$ and for a significance level $\alpha = 0{\cdot}05$, the two-sided test has a power of a little under 0·20 and for $k = 0{\cdot}63$ of a little under 0·93.

must be to give chances of 0·919 and 0·739 of establishing significance at the 10 and 2 % levels respectively, when $p_1 = 0{\cdot}1$, $p_2 = 0{\cdot}6$. From Table 4 we find that $k = 0{\cdot}878$ and interpolating in the second and third columns of Table 5 we obtain h and so n, thus:

P	α	h	$n = h^2/k^2$
0·919	0·10	3·05	12·1 or 13, taking the next higher integer
0·739	0·02	2·97	11·4 or 12, taking the next higher integer

The Table shows the extent of the underestimate of n. The error will become relatively smaller as n is increased; but some adjustment is clearly desirable.

10. A SECOND APPROXIMATION TO OVER-ALL POWER

The first term in the right-hand side of (14) has been taken as the first approximation to the over-all power. The other terms decrease fast since the higher derivatives, $\beta''(\mu_1), \beta^{(4)}(\mu_1), \ldots$ rapidly approach zero, and a good approximation may be obtained by taking a few of these terms. We will, however, derive a second approximation by a slightly different approach.

The method of approximate product-integration developed by R. E. Beard (1947) could be employed to express β as a weighted sum of terms $\beta(r_n)$ where $r_1, r_2, \ldots, r_n$ are n values which might or might not be fixed beforehand. Considering only three such levels of r, we write formally

$$\int_{-\infty}^{\infty} \beta(r)\, p(r)\, dr = \{a_1\beta(r_1) + a_2\beta(r_2) + a_3\beta(r_3)\} \int_{-\infty}^{\infty} p(r)\, dr. \tag{20}$$

Expand the functions $\beta(r)$, $\beta(r_1)$, $\beta(r_2)$ and $\beta(r_3)$ by Taylor's Theorem to six terms as in (13) and write the remainder terms after the sixth. Then identifying the coefficients of

$$\beta(\mu_1), \quad \beta'(\mu_1), \quad \ldots, \quad \beta^{(5)}(\mu_1)$$

on both sides of (20), we have the equations

$$a_1 + a_2 + a_3 = 1$$

$$a_1(r_1 - \mu_1) + a_2(r_2 - \mu_1) + a_3(r_3 - \mu_1) = 0$$

$$a_1(r_1 - \mu_1)^2 + a_2(r_2 - \mu_1)^2 + a_3(r_3 - \mu_1)^2 = \mu_2$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$a_1(r_1 - \mu_1)^5 + a_2(r_2 - \mu_1)^5 + a_3(r_3 - \mu_1)^5 = 0.$$

Expressing the higher moments of the normal distribution $p(r)$ in terms of μ_2, these equations yield the solution: $a_1 = a_3 = \frac{1}{6}$, $a_2 = \frac{2}{3}$.

$$r_1 = \mu_1 - \sqrt{(3\mu_2)}, \quad r_2 = \mu_1, \quad r_3 = \mu_1 + \sqrt{(3\mu_2)}.$$

Since the left-hand side of (20) is β, it follows that

$$\beta = \tfrac{1}{6}\beta[\mu_1 - \sqrt{(3\mu_2)}] + \tfrac{2}{3}\beta(\mu_1) + \tfrac{1}{6}\beta[\mu_1 + \sqrt{(3\mu_2)}] + R, \tag{21}$$

where R depends on the four remainder terms of the expansions $\beta(r), \ldots, \beta(r_3)$. It can be shown that

$$R = \frac{\beta^{(6)}(\xi)}{6!}\left(\mu_6 - \frac{\mu_4^2}{\mu_2}\right) = \frac{\beta^{(6)}(\xi)}{6!}\, 6\mu_2^3 \quad (-\infty < \xi < \infty).$$

We may now compare the expression for β in (21), with that in (14). Expanding the two functions $\beta[\mu_1 - \sqrt{(3\mu_2)}]$ and $\beta[\mu_1 + \sqrt{(3\mu_2)}]$ by Taylor's Theorem, we see that the right-hand side of (21), without R, includes the terms $\beta(\mu_1)$, $\frac{\beta''(\mu_1)}{2!}\mu_2$, $\frac{\beta^{(4)}(\mu_1)}{4!}\mu_4$ of (14) completely

and the following terms partially. This is also seen by comparing the form of R with the form of the remainder term of (14) after the sixth, $R_6 = \beta^{(6)}(\xi)\,\mu_6/6!$. Clearly,

$$\beta = \tfrac{1}{6}\beta[\mu_1 - \sqrt{(3\mu_2)}] + \tfrac{2}{3}\beta(\mu_1) + \tfrac{1}{6}\beta[\mu_1 + \sqrt{(3\mu_2)}] \tag{22}$$

gives a better approximation than

$$\beta = \beta(\mu_1) + \frac{\beta''(\mu_1)}{2!}\mu_2 + \frac{\beta^{(4)}(\mu_1)}{4!}\mu_4.$$

As it is also easier to calculate, we will regard (22) as our second approximation to the over-all power function.

We can get a similar formula with equal weights, namely

$$\beta = \frac{1}{3}\left\{\beta\left[\mu_1 - \frac{\sqrt{(3\mu_2)}}{2}\right] + \beta(\mu_1) + \beta\left[\mu_1 + \frac{\sqrt{(3\mu_2)}}{2}\right]\right\}.$$

Unlike (22), this is derived by using only the first three moments of the distribution of r. More generally, using only the first three moments, we have

$$\beta = \frac{1}{2c^2}\{\beta[\mu_1 - c\sqrt{(\mu_2)}] + \beta[\mu_1 + c\sqrt{(\mu_2)}]\} + \left(1 - \frac{1}{c^2}\right)\beta(\mu_1).$$

where c can be chosen as we like, subject to the restriction that none of the arguments should become negative.

In section 8 we have seen how the over-all power for equal samples could be obtained to the first approximation, $\beta(\mu_1)$. For the second approximation (22) the values of $\beta[\mu_1 + \sqrt{(3\mu_2)}]$ and $\beta[\mu_1 - \sqrt{(3\mu_2)}]$ may be obtained in the same manner by entering Table 5 with the values of $h[\mu_1 + \sqrt{(3\mu_2)}]$ and $h[\mu_1 - \sqrt{(3\mu_2)}]$. From (6), with $m = n$, we have

$$h[\mu_1 + \sqrt{(3\mu_2)}]$$
$$= \sqrt{\frac{1}{8n}\{(n(p_1+p_2) + \sqrt{[3n(p_1q_1+p_2q_2)]})\,(2n - np_1 + p_2) - \sqrt{[3n(p_1q_1+p_2q_2)]}\}\log_e\frac{p_1q_2}{p_2q_1}}.$$

Putting $h_1 = |\,h[\mu_1 + \sqrt{(3\mu_2)}]\,|$, this becomes

$$h_1 = k\sqrt{\left\{1 + \frac{2(1-p_1-p_2)\sqrt{[3n(p_1q_1+p_2q_2)]} - 3n(p_1q_1+p_2q_2)}{n(p_1+p_2)\,(2-p_1-p_2)}\right\}},$$

where k is the expression (17). Similarly if

$$h_2 = |\,h[\mu_1 - \sqrt{(3\mu_2)}]\,|,$$

then
$$h_2 = k\sqrt{\left\{1 + \frac{-2(1-p_1-p_2)\sqrt{[3n(p_1q_1+p_2q_2)]} - 3n(p_1q_1+p_2q_2)}{n(p_1+p_2)\,(2-p_1-p_2)}\right\}}.$$

h_1 and h_2 can be calculated by obtaining k from Table 4 and substituting the values of k, p_1, p_2 and n in these two expressions.

If P_1, P_2 are the tabulated values of the integral P of (19), corresponding to h_1, h_2 and if P is the value corresponding to $h(\mu_1)$, then the second approximation is

$$\beta = \tfrac{1}{6}(P_1 + P_2 + 4P). \tag{23}$$

The value of n obtained from Table 5 as described in section 8, may be improved with the help of the second approximation. Taking this value of n we calculate h_1 and h_2 and obtain the corresponding P_1 and P_2. Then from (23),

$$P = \tfrac{3}{2}\beta - \tfrac{1}{4}(P_1 + P_2).$$

With this P we enter Table 5 again and find the improved n.

[*From* Biometrika, *Vol.* 46. *Parts* 3 *and* 4, *December* 1959.]
[*All rights reserved.*] PRINTED IN GREAT BRITAIN

CORRIGENDA

(1) *Biometrika* (1948), **35**, pp. 157–75

'The power function of the test for the difference between two proportions in a 2×2 table.'
By P. B. PATNAIK.

I am indebted to Mr R. L. Turner for pointing out some errors in formulae on pp. 172–3 of this paper:

(*a*) p. 172, the expressions for h_1 and h_2 should read

$$h_1 = k\sqrt{\left\{n + \frac{2(1-p_1-p_2)\sqrt{[3n(p_1q_1+p_2q_2)]} - 3(p_1q_1+p_2q_2)}{(p_1+p_2)(2-p_1-p_2)}\right\}},$$

$$h_2 = k\sqrt{\left\{n + \frac{-2(1-p_1-p_2)\sqrt{[3n(p_1q_1+p_2q_2)]} - 3(p_1q_1+p_2q_2)}{(p_1+p_2)(2-p_1-p_2)}\right\}}.$$

(*b*) p. 173, in equation (25) *for* $(1+\theta\overline{1-p_1-p_2})$ *read* $(1+\theta(1-p_1)-p_2)$, an ex-

as the numerical values for the second approximation to the power function given in the examples in Table 3 have been found to be correct.

P. B. PATNAIK

The values of the over-all power have been calculated on this second approximation for certain cases and shown alongside the first approximation values in Tables 3(*a*) and 3(*b*). It is seen that the second approximation improves the first, although it does not remove all the error.

11. General case of unequal sample sizes

The first approximation to the over-all power depends on

$$h(\mu_1) = \sqrt{\left\{\frac{mn(mp_1+np_2)(N-mp_1-np_2)}{N^3}\right\}}\log_e\frac{p_1q_2}{p_2q_1}, \tag{24}$$

taking N^3 instead of $N^2(N-1)$ in (6).

Suppose $\theta = m/n$ then

$$h = |h(\mu_1)| = \sqrt{\frac{N}{2}}\sqrt{\left\{\frac{2\theta}{(1+\theta)^4}(\theta p_1+p_2)(1+\theta\overline{1-p_1-p_2})\right\}}\left|\log_e\frac{p_1q_2}{p_2q_1}\right|$$
$$= \sqrt{(\tfrac{1}{2}N)}\,k(\theta), \quad \text{say.} \tag{25}$$

$k(\theta)$ corresponds to k in the case of equal samples, but in general it is a function of θ as well as p_1 and p_2. If θ is known, as for example, when m and n are given, the k function defines as before a family of contours of constant power in the p_1p_2 plane.

To obtain a first approximation to the power, we evaluate $h(\mu_1)$ from (24) or (25) and enter Table 5. For the second approximation we have to evaluate also $h[\mu_1 \pm \sqrt{(3\mu_2)}]$ and as before obtain the weighted sum of the corresponding P's.

The converse problem of determining the sample sizes for a given power can be solved only if the value of θ is given. For example, we may ask, 'What sample sizes should we take, one being double the other if we want a 90 % chance of detecting a significant difference when p_1 and p_2 have certain specified values?'

First we calculate $k(\theta)$. From the given power we obtain h from Table (5). Then $N = k^2(\theta)/h^2$ will give N to the first approximation. Since θ is given, m and n are obtained.

12. Summary

The power function of the test of significance for the 2×2 table has been considered and an approximate method of deriving it has been developed. The usefulness of the idea of power in fixing in advance the sizes of samples is indicated.

Tables have been provided for determining the power for a specified alternative, the samples sizes being given and conversely.

I wish to express my grateful thanks to Prof. E. S. Pearson and Dr H. O. Hartley for their guidance in the study of this problem.

REFERENCES

BARNARD, G. A. (1947). *Biometrika*, **34**, 123.
BEARD, R. E. (1947). *J. Inst. Actu.* **73**, 356.
NEYMAN, J. & PEARSON, E. S. (1933). *Philos. Trans.* A, **231**, 289.
NEYMAN, J. & TOKARSKA (1936). *J. Amer. Statist. Ass.* **31**, 318.
PEARSON, E. S. (1947). *Biometrika*, **34**, 139.
TANG, P. C. (1938). *Statist. Res. Mem.* **2**, 126.

Table 4. *Values* of $k = \frac{\sqrt{2}}{4}\sqrt{\{(p_1+p_2)(2-p_1-p_2)\}}\log_e\frac{p_1(1-p_2)}{p_2(1-p_1)}$

[For $p_1 < p_2$, enter with $1-p_1$ and $1-p_2$]

p_2	p_1 = 0	0·05	0·10	0·15	0·20	0·25	0·30	0·35	0·40	0·45	0·50	0·55	0·60	0·65	0·70	0·75	0·80	0·85	0·90	0·95	1·00
1·00																					0
0·95																				0	0·139
0·90																			0	0·139	0·257
0·85																		0	0·107	0·257	0·359
0·80																	0	0·094	0·205	0·359	0·466
0·75																0	0·085	0·180	0·295	0·466	0·563
0·70															0	0·079	0·165	0·262	0·382	0·563	0·658
0·65														0	0·076	0·155	0·242	0·342	0·466	0·658	0·750
0·60													0	0·073	0·149	0·230	0·318	0·420	0·549	0·750	0·840
0·55												0	0·072	0·145	0·221	0·303	0·393	0·497	0·630	0·840	0·930
0·50											0	0·071	0·143	0·216	0·294	0·376	0·468	0·574	0·712	0·930	1·019
0·45										0	0·071	0·142	0·214	0·288	0·366	0·450	0·543	0·653	0·794	1·019	1·109
0·40									0	0·072	0·143	0·214	0·287	0·362	0·441	0·526	0·621	0·733	0·878	1·109	1·202
0·35								0	0·073	0·145	0·216	0·288	0·362	0·438	0·518	0·604	0·714	0·815	0·964	1·202	1·298
0·30							0	0·076	0·149	0·221	0·294	0·366	0·441	0·518	0·599	0·687	0·786	0·903	1·055	1·298	1·401
0·25						0	0·079	0·155	0·230	0·303	0·376	0·450	0·526	0·604	0·687	0·777	0·877	0·997	1·152	1·401	1·514
0·20					0	0·085	0·165	0·242	0·318	0·393	0·468	0·543	0·621	0·714	0·786	0·877	0·980	1·102	1·261	1·514	1·646
0·15				0	0·094	0·180	0·262	0·342	0·420	0·497	0·574	0·653	0·733	0·815	0·903	0·997	1·102	1·227	1·388	1·646	1·816
0·10			0	0·107	0·205	0·295	0·382	0·466	0·549	0·630	0·712	0·794	0·878	0·964	1·055	1·152	1·261	1·388	1·554	1·816	2·082
0·05		0	0·139	0·257	0·359	0·466	0·563	0·658	0·750	0·840	0·930	1·019	1·109	1·202	1·298	1·401	1·514	1·646	1·816	2·082	
0																					0

Values of p_2

Values of p_1

Table 5. *Relating Power and k, h, n and α*

P of equation (19) or the approximate power					Values of $k = h/\sqrt{n}$, according to h and n															
h \ α	0·10	0·05	0·02	0·01	10	15	20	25	30	35	40	45	50	60	70	80	90	100	150	h \ n
0·1	0·1017	0·0511	0·0206	0·0104*	0·032	0·026	0·022	0·020	0·018	0·017	0·016	0·015	0·014	0·013	0·012	0·011	0·011	0·010	0·008	0·1
0·2	0·1068	0·0546	0·0225	0·0115	0·063	0·052	0·045	0·040	0·037	0·034	0·032	0·030	0·028	0·026	0·024	0·022	0·021	0·020	0·016	0·2
0·3	0·1152	0·0604	0·0257	0·0134	0·095	0·077	0·067	0·060	0·055	0·051	0·047	0·045	0·042	0·039	0·036	0·034	0·032	0·030	0·024	0·3
0·4	0·1270	0·0685	0·0302	0·0162	0·126	0·103	0·089	0·080	0·073	0·068	0·063	0·060	0·057	0·052	0·048	0·045	0·042	0·040	0·033	0·4
0·5	0·1421	0·0791	0·0363*	0·0200	0·158	0·129	0·112	0·100	0·091	0·085	0·079	0·075	0·071	0·065	0·060	0·056	0·053	0·050	0·041	0·5
0·6	0·1604	0·0922	0·0439	0·0248	0·190	0·155	0·134	0·120	0·110	0·101	0·095	0·089	0·085	0·077	0·072	0·067	0·063	0·060	0·049	0·6
0·7	0·1819	0·1077	0·0532	0·0309	0·221	0·181	0·157	0·140	0·128	0·118	0·111	0·104	0·099	0·090	0·084	0·078	0·074	0·070	0·057	0·7
0·8	0·2063	0·1259	0·0643	0·0382	0·253	0·207	0·179	0·160	0·146	0·135	0·126	0·119	0·113	0·103	0·096	0·089	0·084	0·080	0·065	0·8
0·9	0·2336	0·1467	0·0775	0·0471	0·285	0·232	0·201	0·180	0·164	0·152	0·142	0·134	0·127	0·116	0·108	0·101	0·095	0·090	0·073	0·9
1·0	0·2636	0·1701	0·0928	0·0577	0·316	0·258	0·224	0·200	0·183	0·169	0·158	0·149	0·141	0·129	0·120	0·112	0·105	0·100	0·082	1·0
1·1	0·2960*	0·1960*	0·1103	0·0701	0·348	0·284	0·246	0·220	0·201	0·186	0·174	0·164	0·156	0·142	0·131	0·123	0·116	0·110	0·090	1·1
1·2	0·3304	0·2244	0·1302	0·0845†	0·379	0·310	0·268	0·240	0·219	0·203	0·190	0·179	0·170	0·155	0·143	0·134	0·126	0·120	0·098	1·2
1·3	0·3667	0·2552	0·1525†	0·1011	0·411	0·336	0·291	0·260	0·237	0·220	0·206	0·194	0·184	0·168	0·155	0·145	0·137	0·130	0·106	1·3
1·4	0·4044	0·2881	0·1772	0·1199‡	0·443	0·361	0·313	0·280	0·256	0·237	0·221	0·209	0·198	0·181	0·167	0·157	0·148	0·140	0·114	1·4
1·5	0·4432	0·3230	0·2044	0·1410	0·474	0·387	0·335	0·300	0·274	0·254	0·237	0·224	0·212	0·194	0·179	0·168	0·158	0·150	0·122	1·5
1·6	0·4827	0·3596	0·2339	0·1646	0·506	0·413	0·358	0·320	0·292	0·270	0·253	0·239	0·226	0·207	0·191	0·179	0·169	0·160	0·131	1·6
1·7	0·5224	0·3976	0·2656‡	0·1906	0·538	0·439	0·380	0·340	0·310	0·287	0·269	0·253	0·240	0·219	0·203	0·190	0·179	0·170	0·139	1·7
1·8	0·5619	0·4365†	0·2993	0·2189	0·569	0·465	0·402	0·360	0·329	0·304	0·285	0·268	0·255	0·232	0·215	0·201	0·190	0·180	0·147	1·8
1·9	0·6009	0·4761	0·3349	0·2496	0·601	0·491	0·425	0·380	0·347	0·321	0·300	0·283	0·269	0·245	0·227	0·212	0·200	0·190	0·155	1·9
2·0	0·6389	0·5160	0·3721	0·2824	0·632	0·516	0·447	0·400	0·365	0·338	0·316	0·298	0·283	0·258	0·239	0·224	0·211	0·200	0·163	2·0
2·1	0·6756†	0·5557	0·4105	0·3171	0·664	0·542	0·470	0·420	0·383	0·355	0·332	0·313	0·297	0·271	0·251	0·235	0·221	0·210	0·171	2·1
2·2	0·7107	0·5949	0·4497	0·3535	0·696	0·568	0·492	0·440	0·402	0·372	0·348	0·328	0·311	0·284	0·263	0·246	0·232	0·220	0·180	2·2
2·3	0·7439	0·6331	0·4895	0·3913	0·727	0·594	0·514	0·460	0·420	0·389	0·364	0·343	0·325	0·297	0·275	0·257	0·242	0·230	0·188	2·3
2·4	0·7749	0·6701‡	0·5294	0·4302	0·759	0·620	0·537	0·480	0·438	0·406	0·379	0·358	0·339	0·310	0·287	0·268	0·253	0·240	0·196	2·4
2·5	0·8038	0·7054	0·5689	0·4698	0·791	0·645	0·559	0·500	0·456	0·423	0·395	0·373	0·354	0·323	0·299	0·280	0·264	0·250	0·204	2·5
2·6	0·8303‡	0·7389	0·6078	0·5096	0·822	0·671	0·581	0·520	0·475	0·439	0·411	0·388	0·368	0·336	0·311	0·291	0·274	0·260	0·212	2·6
2·7	0·8543	0·7704	0·6457	0·5494	0·854	0·697	0·604	0·540	0·493	0·456	0·427	0·402	0·382	0·349	0·323	0·302	0·285	0·270	0·220	2·7
2·8	0·8760	0·7996	0·6821	0·5887	0·885	0·723	0·626	0·560	0·511	0·473	0·443	0·417	0·396	0·361	0·335	0·313	0·295	0·280	0·229	2·8
2·9	0·8953	0·8264	0·7169	0·6271	0·917	0·749	0·648	0·580	0·529	0·490	0·459	0·432	0·410	0·374	0·347	0·324	0·306	0·290	0·237	2·9
3·0	0·9123	0·8508	0·7497	0·6643	0·949	0·775	0·671	0·600	0·548	0·507	0·474	0·447	0·424	0·387	0·359	0·335	0·316	0·300	0·245	3·0
3·2	0·9400	0·8925	0·8088	0·7337	1·012	0·826	0·716	0·640	0·584	0·541	0·506	0·477	0·453	0·413	0·382	0·358	0·337	0·320	0·261	3·2
3·4	0·9604	0·9251	0·8585	0·7951	1·075	0·878	0·760	0·680	0·621	0·575	0·538	0·507	0·481	0·439	0·406	0·380	0·337	0·340	0·278	3·4
3·6	0·9747	0·9495	0·8986	0·8471	1·138	0·930	0·805	0·720	0·657	0·609	0·569	0·537	0·509	0·465	0·430	0·402	0·348	0·360	0·294	3·6
3·8	0·9844	0·9671	0·9297	0·8896	1·202	0·981	0·850	0·760	0·694	0·642	0·601	0·566	0·537	0·491	0·454	0·425	0·358	0·380	0·301	3·8
4·0	0·9907	0·9793	0·9529	0·9228	1·265	1·033	0·894	0·800	0·730	0·676	0·632	0·596	0·566	0·516	0·478	0·447	0·369	0·400	0·327	4·0
4·2	0·9947	0·9875	0·9695	0·9478	1·328	1·084	0·939	0·840	0·767	0·710	0·664	0·626	0·594	0·542	0·502	0·470	0·379	0·420	0·343	4·2
4·4	0·9971	0·9927	0·9809	0·9659	1·391	1·136	0·984	0·880	0·803	0·744	0·696	0·656	0·622	0·568	0·526	0·492	0·390	0·440	0·359	4·4
4·6	0·9984	0·9959	0·9885	0·9785	1·455	1·188	1·029	0·920	0·840	0·778	0·727	0·686	0·651	0·594	0·550	0·514	0·401	0·460	0·376	4·6
4·8	0·9992	0·9977	0·9933	0·9869	1·518	1·239	1·073	0·960	0·876	0·811	0·759	0·716	0·679	0·620	0·574	0·537	0·411	0·480	0·392	4·8
5·0	0·9996	0·9988	0·9962	0·9923	1·581	1·291	1·118	1·000	0·913	0·845	0·791	0·745	0·707	0·645	0·598	0·559	0·422	0·500	0·408	5·0

THE ANALYSIS OF CONTINGENCY TABLES WITH GROUPINGS BASED ON QUANTITATIVE CHARACTERS

By F. YATES

A $p \times q$ contingency table can be tested for independence by a χ^2 test with $(p-1)(q-1)$ degrees of freedom. This is an over-all test which covers all forms of departure from proportionality, and is consequently correspondingly insensitive to departures of a specified type.

If the nature of the data is such that departures of a particular type are to be expected, then a test of significance appropriate to departures of this type will be justified.

The present paper deals with the case in which one or both groupings are based on characters which are either directly quantitative or are in the form of gradings which can be regarded as having an underlying quantitative basis. The actual data which gave rise to the investigation are shown in Table 1. They were obtained in the course of a pilot inquiry into the conditions in which school children do their homework, carried out by the Department of Social Science, University of Liverpool, and I am indebted to Mr D. Chapman for permission to reproduce them here.

Table 1. *Relation (in terms of numbers of children and percentages) between conditions under which homework was carried out, and the teacher's rating of the quality of that homework. (Each scale is graded, A being the highest rating)*

Teacher's rating	Homework conditions					
	A	*B*	*C*	*D*	*E*	Total
A	141 (46%)	67 (46%)	114 (39%)	79 (44%)	39 (43%)	440 (43%)
B	131 (42%)	66 (45%)	143 (48%)	72 (40%)	35 (39%)	447 (44%)
C	36 (12%)	14 (9%)	38 (13%)	28 (16%)	16 (18%)	132 (13%)
Total	308 (100%)	147 (100%)	295 (100%)	179 (100%)	90 (100%)	1019 (100%)

It is clear from the percentages that the effect of homework conditions on the quality of the preparation, as judged by the teacher's rating, is small. On the other hand, there is some slight trend, and the question therefore arises whether this trend has any significance, or, more generally, what is its estimated magnitude, and what are the errors of this estimate.

A χ^2 test of the whole table gives a value of χ^2 equal to 9·16 (8 degrees of freedom), but such a test, as pointed out above, embraces all types of deviation from proportionality.

In material of this kind a rough test, which isolates the major part of any quantitative association between the two variates, can be made by reduction of the table to a 2×2 table by the grouping of appropriate parts and rejection of others. With the above data we might reasonably group homework conditions *A* and *B*, and *D* and *E*, rejecting condition *C*, and also reject the teacher's central rating *B*. This will give the values of Table 2.

This Table gives a value of χ_c^2, i.e. χ^2 corrected for continuity, equal to 3·03 (1 degree of freedom), corresponding to a probability (for one tail of the χ_c distribution) of 0·041. (The use of a single tail is appropriate to testing whether there is evidence that improved conditions of homework improve the quality of preparation, ruling out the opposite contingency.) The test therefore indicates that there is significant evidence (at the 5 % level) of some improvement.

The above rough test, however, is open to several objections. In the first place, since the grouping is arbitrary, there is always a possibility that the statistician will allow himself to be influenced by the data in his choice of grouping, a grouping which gives a high degree of association being chosen. If this occurs, the test of significance is clearly vitiated. Secondly, different workers may in good faith choose different groupings of the same data, and this may lead to arguments of a type that are likely to discredit the science of statistics. Thirdly, the choice of grouping which is most appropriate in any given case depends on the marginal totals of the numbers of observations, and simple rules for the choice of grouping are not easy to devise. Fourthly, the test provides no estimate of the magnitude of the effects of the association.

Table 2. *Condensation of part of Table* 1

Teacher's rating	Homework conditions		
	A+*B*	*D*+*E*	Total
A	208 (81 %)	118 (73 %)	326 (78 %)
C	50 (19 %)	44 (27 %)	94 (22 %)
Total	258 (100 %)	162 (100 %)	420 (100 %)

The test based on regression concepts, developed below, eliminates the element of choice, requires no elaborate computation, and also provides estimates of the magnitude of the effect of each variate on the other.

The following notation will be adopted. Letters without dashes will be taken to indicate the results of operations on the rows, and with dashes the results of the same operations on the columns. The r row totals will be denoted by $N_1, N_2, \ldots$, the r' column totals by $N'_1, N'_2, \ldots$, and the grand total by T. We shall also require an extended summation notation, in which the x's may denote any set of r quantities, as follows,

$$S_0(x) = x_1 + x_2 + \ldots + x_r,$$
$$S_1(x) = -\tfrac{1}{2}(r-1)x_1 - \tfrac{1}{2}(r-3)x_2 - \ldots + \tfrac{1}{2}(r-1)x_r,$$
$$S_2(x) = \tfrac{1}{4}(r-1)^2 x_1 + \tfrac{1}{4}(r-3)^2 x_2 + \ldots + \tfrac{1}{4}(r-1)^2 x_r,$$

with three similar functions extending to r' terms distinguished by dashes.

For each of the r' columns a mean score is calculated, assigning a score of $-\frac{1}{2}(r-1)$ to the first row, $-\frac{1}{2}(r-3)$ to the second row, etc. and $+\frac{1}{2}(r-1)$ to the last row. If the numbers in the different sub-groups of the first column are $n_1, n_2, \ldots, n_r$, the mean score for the column is

$$u'_1 = \frac{1}{N'_1} S_1(n),$$

and the total score is

$$U'_1 = S_1(n).$$

If the numbers $n_1, n_2, \ldots, n_r$ are regarded as a sample from a multinomial distribution with probabilities of, say, $p_1, p_2, \ldots, p_r$, the variance of n_1 will be $N_1' p_1 q_1$, and the covariance of n_1 and n_2 will be $-N_1' p_1 p_2$, etc. The variance of u_1' can then be simply calculated, and will be found to be

$$V(u_1') = \frac{1}{N_1'}[S_2(p) - \{S_1(p)\}^2].$$

The regression of the mean scores $u_1', u_2', \ldots, u_{r'}'$ for the r' columns on column number can now be calculated, weighting according to the variances of the u''s, provided we can make some assumption as to the values of the p's.

In order to obtain a test of significance we may start with the hypothesis that the sets of p's are identical for all columns, i.e. that p_1 is the same from column to column, etc. Then the variances of the u''s are inversely proportional to the numbers in the columns, i.e. the N''s, and in calculating the regression we may therefore take weights equal to the N''s. Estimates of the values of p's may be derived from the row totals, so that $p_1 = N_1/T$, etc., when we shall have

$$V(u_1') = \frac{1}{N_1' T^2}[T S_2(N) - \{S_1(N)\}^2],$$

etc.

If the regression equation is taken in the form

$$u_x' = m' + b'\{x - \tfrac{1}{2}(r+1)\},$$

the equations of estimation for m' and b' are:

$$m' S_0'(N') + b' S_1'(N') = S_0'(U'), \quad m' S_1'(N') + b' S_2'(N') = S_1'(U').$$

We have

$$S_0(N) = S_0'(N') = T, \quad S_0'(U') = S_1(N),$$
$$S_0(U) = S_1'(N'), \qquad S_1(U) = S_1'(U').$$

If we put

$$A = T S_2(N) - \{S_1(N)\}^2,$$
$$A' = T S_2'(N') - \{S_1'(N')\}^2,$$
$$B = T S_1(U) - S_1(N) S_1'(N'),$$

the solution of the above equations gives $b' = B/A'$, with the corresponding regression on row scores $b = B/A$.

The variance per unit weight is A/T^2, and consequently

$$V(b') = \frac{T}{A'} \frac{A}{T^2} = \frac{A}{A'T}.$$

Similarly, $V(b) = A'/AT$. The appropriate test of significance is therefore given by

$$\chi^2 = \frac{b^2}{V(b)} = \frac{b'^2}{V(b')} = \frac{B^2 T}{AA'}$$

with one degree of freedom.

The above test is unaffected by interchange of rows and columns, as should be the case. It can easily be verified that the test reduces to the ordinary χ^2 test of a 2×2 contingency table (apart from the correction for continuity) if the values in all but two of the rows and two of the columns are put equal to zero. The correction for continuity is not here of importance in view of the large number of possible alternatives with given marginal totals.

The test has been arrived at by assuming that the column totals only are fixed, but with the additional approximation involved in assuming that the p's for each column are given by the marginal totals for the rows. This approach has the advantage of indicating the

appropriate criterion B^2T/AA' for the type of association which it is desired to test. Once the criterion is determined, however, it can be shown, following Fisher (1922, 1925), that this criterion, which is linear in the deviations from expectation, and orthogonal with the linear functions representing the row and column totals, will in large samples be distributed as χ^2 with 1 degree of freedom when both sets of marginal totals are held fixed.

The values for the regression coefficients b and b' give estimates of the change in mean score with unit change of row and column respectively. It should be noted that the variances of b and b' given above are based on the assumption that there is no association between the two variates and become progressively more inaccurate as the degree of association increases. In general they will be over-estimates of the true variances. Nothing more accurate is likely to be required, however, except when there is very marked association, in which case the assumption of linear regression in the mean scores over the whole range is unlikely to provide an adequate mathematical description of the association.

If only one of the classifications, say the rows, is quantitative, a test for the homogeneity of the mean scores for the columns may be derived from the variances of these mean scores. The quantity

$$\begin{aligned} Q &= S_0'\left(\frac{u'^2}{V(u')}\right) - \frac{[S_0'\{u'/V(u')\}]^2}{S_0'\{1/V(u')\}} \\ &= \frac{T^2}{A}\left[S_0'\left(\frac{U'^2}{N'}\right) - \frac{\{S_0'(U')\}^2}{T}\right] \\ &= \frac{T^2}{A}[S_0'(u'U') - \bar{u}'S_0'(U')], \end{aligned}$$

where $\bar{u}'$ is the mean score for all the observations, will be distributed as χ^2 with $r'-1$ degrees of freedom. It is easily verified that this test reduces to the ordinary χ^2 test for a $2 \times r'$ table when there are only two rows.

It may be noted that any system of scoring may be assigned to each of the classifications—there is no need to adopt a system with equal intervals between each class if the nature of the classification is such that scoring with unequal intervals is more appropriate. If, for example, in the data of Table 1 the teachers had given their opinion that the difference between their classes A and B in quality of preparation was only half that between B and C scores of $+1, 0$, and -2 might have been adopted. In the absence of any such indications, however, the appropriate procedure will be to assume that graded classifications are intended to represent equal intervals on some scale, unless the data themselves are used to determine the optimal system of scoring, by some procedure analogous to that given by Fisher (1946), para. 49·2, or by reference to some assumed distribution of the scores in the population. Examples of the latter procedure are described by Pearson & Moul (1925), and also in K. Pearson's *Tables for Statisticians and Biometricians*, Part II, pp. xxiii–xxvi (1931). In this connexion see also E. S. Pearson (1923). Such procedures, however, are only likely to be worth while in exceptional cases, and with very extensive data.

The computational procedure when the scores are assumed is very simple, and is illustrated in Table 3 for a 4×5 table with quantitative classifications. The scores used for the rows have been multiplied by 2 for convenience in computation. The value of b obtained will therefore represent half the change to be expected in the mean score from row to row.

The total scores U and U' in Table 3 are calculated by summing the products of the scores and the numbers in the corresponding sub-classes. Checks are provided by carrying out the same operations on the totals. $S_2(N)$ and $S_2(N')$ are obtained by multiplying the totals

$N_1, N_2, \ldots$, and $N_1', N_2', \ldots$, by the squares of the scores. These must be checked. $S_1(U)$ is obtained in two ways by summing the products of the U's and the U''s with their corresponding scores.

The quantity A is then obtained from the last three values in the total column, A' from the last three values in the total line, and B from the cross product of the 2×2 table formed by the total and total score rows and columns.

Table 3. *Computational procedure*

		(Score)²	4	1	0	1	4		
		Score	−2	−1	0	+1	+2		
(Score)²	Score		*A*	*B*	*C*	*D*	*E*	Total	Total score
9	−3	*A*	—	—	—	—	—	N_1	U_1
1	−1	*B*	—	—	—	—	—	N_2	U_2
1	+1	*C*	—	—	—	—	—	N_3	U_3
9	+3	*D*	—	—	—	—	—	N_4	U_4
		Total	N_1'	N_2'	N_3'	N_4'	N_5'	T	$S_0(U) = S_1'(N')$ $S_2(N') \leftarrow A'$
		Total score	U_1'	U_2'	U_3'	U_4'	U_5'	$S_0'(U') = S_1(N)$	$S_1(U) = S_1'(U')$ ↖ B
								$S_2(N)$ ↑ A	

Table 4. *Analysis of the data of Table* 1

			Homework conditions								
		(Score)²	4	1	0	1	4				
		Score	+2	+1	0	−1	−2				
(Score)²	Score	Teacher's rating	*A*	*B*	*C*	*D*	*E*	Total	Total score	Total (score)²	Mean score
1	+1	*A*	141	67	114	79	39	440	+192		+0·44
0	0	*B*	131	66	143	72	35	447	+186		+0·42
1	−1	*C*	36	14	38	28	16	132	+ 26		+0·20
		Total	308	147	295	179	90	1019	+404	1918	+0·40
		Total score	+105	+53	+76	+51	+23	+308	+166		
		Total (score)²						572			
		Mean score	+0·34	+0·36	+0·26	+0·28	+0·26	+0·30			

The analysis of the data of Table 1 is given in Table 4, which also shows the mean scores (not required in the analysis). From the values of this table we have:

$$A = 1019 \times 572 - 308^2 = 488{,}004,$$

$$A' = 1019 \times 1918 - 404^2 = 1{,}791{,}226,$$

$$B = 1019 \times 166 - 308 \times 404 = 44{,}722,$$

$$b = B/A = 44{,}722/488{,}004 = 0{\cdot}09164,$$

$$b' = B/A' = 44{,}722/1{,}791{,}226 = 0{\cdot}02497,$$

$$\text{S.E. of } b = \sqrt{\frac{A'}{AT}} = \sqrt{\frac{1{,}791{,}226}{488{,}004 \times 1019}} = \pm 0{\cdot}0600,$$

$$\text{S.E. of } b' = \sqrt{\frac{A}{A'T}} = \sqrt{\frac{488{,}004}{1{,}791{,}226 \times 1019}} = \pm 0{\cdot}0163,$$

$$\chi^2 = \frac{B^2T}{AA'} = \frac{44{,}722^2 \times 1019}{488{,}004 \times 1{,}791{,}226} = 2{\cdot}332.$$

Reference to a table of the normal integral indicates that $P = 0{\cdot}062$ (one tail), and there is therefore a probability about 1 in 16 of obtaining by chance as great an apparent improvement as is indicated by the data, if homework conditions have in fact no influence on the quality of the homework as shown by the teacher's rating. More important, in data of this kind, the analysis indicates that the amount of improvement, if any, is likely to be small. In terms of the improvement to be expected in changing conditions from the worst (E) to the best (A) the estimated improvement, $4b'$, is 0·100, the limits given by once and twice the standard error being $+0{\cdot}03$ to $+0{\cdot}16$ and $-0{\cdot}03$ to $+0{\cdot}23$. We may therefore state that the true value of this improvement is not likely to be substantially greater than $+0{\cdot}16$ and is almost certainly not greater than $+0{\cdot}23$, nor is it likely to be substantially less than $+0{\cdot}03$, and is almost certainly not less than $-0{\cdot}03$.

It will be noted that the exact test in this instance gives a lower degree of significance than the rough test based on reduction to a 2×2 table. This, of course, does not imply that the exact test is less sensitive. With any given set of data the verdicts of different tests of significance may differ considerably owing to chance causes. This does not preclude the use of rough tests when considerations of speed and computational labour are paramount, but once the exact test has been evaluated the verdict of the simpler but less appropriate test must be set aside. The decision as to the test to be used should also be made without reference to the actual data—it is, of course, inadmissible to make a rough test, accept its verdict if significant, and proceed to an exact test if it fails to indicate significance.

As an example of the computations when one classification only is quantitative we may test the deviations between the mean scores for different homework conditions. This test would be required if the homework conditions were qualitative categories and not an ordered series. In this case the value of χ^2 for 4 degrees of freedom will be given by

$$Q = \frac{1019^2}{488{,}004}(105 \times 0{\cdot}34091 + \ldots - 308 \times 0{\cdot}302257)$$

$$= 3{\cdot}825,$$

the mean scores being taken to 5 and 6 decimal places to ensure the necessary accuracy. It is clear that apart from the linear trend there are no significant variations in mean score.

REFERENCES

FISHER, R. A. (1922). *J. R. Statist. Soc.* **85**, 87.

FISHER, R. A. (1925). *Metron*, **5**, 3.

FISHER, R. A. (1946). *Statistical Methods for Research Workers*, 10th ed. Edinburgh: Oliver and Boyd.

PEARSON, E. S. (1923). *Biometrika*, **14**, 261.

PEARSON, K. (1931). *Tables for Statisticians and Biometricians*, Part II. London: Biometrika Office.

PEARSON, K. & MOUL, M. (1925). *Ann. Eugen., Lond.*, **1**, 1.

THE PROBABILITY INTEGRAL TRANSFORMATION WHEN PARAMETERS ARE ESTIMATED FROM THE SAMPLE

By F. N. DAVID and N. L. JOHNSON

1. The probability integral transformation for testing goodness of fit and combining tests of significance was introduced by R. A. Fisher in 1932. Fisher's objective was the significance of combined independent tests of significance, but his method also proved applicable to a certain limited range of tests for goodness of fit as can be seen in K. Pearson (1933), J. Neyman (1937) and E. S. Pearson (1938). The transformation may be summarized briefly in the following way. Assume that there is a continuous random variable x whose elementary probability law is $p(x)$, whence obviously

$$\int_{-\infty}^{+\infty} p(x)\,dx = 1.$$

Consider a new random variable, y, connected with x by the relation

$$y = \int_{-\infty}^{x} p(x)\,dx.$$

y is a monotonic non-decreasing function of x and $0 \leqslant y \leqslant 1$. Further, if $p(y)$ is the elementary probability law of y, then

$$p(y) = p(x)\frac{dx}{dy} = 1.$$

Hence in the interval $[0; 1]$ all values of y are equally likely, or in common parlance, y is rectangularly distributed in the interval $[0; 1]$, no matter what the elementary probability law of x. If therefore we have n independent random variables $x_j (j = 1, 2, \ldots, n)$ following a known continuous probability law which is completely specified by H_0, the hypothesis tested, then by means of the transformation

$$y_j = \int_{-\infty}^{x_j} p(x_j \mid H_0)\,dx_j,$$

the x's can be transformed into n independent random variables y which are rectangularly distributed.

2. The transformation which we have just summarized is useful statistically in that tests based on a rectangular population can be made applicable to any variable of which the elementary probability law is known. However, because the parameters of the elementary probability law must be specified, it is clear that the range of application of any tests based on this transformation will be very restricted, for cases are rare in statistical practice when H_0 is completely specified. It seemed interesting to us to investigate the effect on the transformation of calculating estimates of the parameters from the data provided by the sample. For example, if the mean of the probability law is estimated from a sample of n quantities $X_1, X_2, \ldots, X_n$ each of which is one observed value of n random variables $x_1, x_2, \ldots, x_n$, the y's obtained by the probability integral transformation will no longer be independent, neither will they be rectangularly distributed. We are able to show that the generality of the transformation in the case when the parameters are completely specified is lost as soon as we begin replacing unknown parameters by the sample estimates, and, as is intuitively obvious, the form of the probability law of y depends on the functional form of the common probability law of the x's.

3. We begin by stating the problem in a formal mathematical way and indicate the method whereby a general solution is reached. Assume that $p(x)$ is a single valued continuous function of the form

$$p(x) = f(x \mid \theta_1, \theta_2, \ldots, \theta_s),$$

where, in the usual way, $\theta_1, \theta_2, \ldots, \theta_s$, are parameters descriptive of the population, all of which may or may not be specified. It may be assumed for generality that none are specified and that in place of the unknown parameters, θ, it is necessary to substitute functions of the sample values, say,

$$F_1(x_1, x_2, \ldots, x_n), \quad F_2(x_1, x_2, \ldots, x_n), \quad \ldots, \quad F_s(x_1, x_2, \ldots, x_n),$$

where $x_1, x_2, \ldots, x_n$ are the random variables of which the n observations which form the sample are the observed values. Thus we require to find the distribution of the variables

$$y_i = \int_{-\infty}^{x_i} f(t \mid F_1, F_2, \ldots, F_s)\, dt \quad \text{for} \quad i = 1, 2, \ldots, n.$$

We have immediately that

$$\frac{\partial y_i}{\partial x_j} = \int_{-\infty}^{x_i} \sum_{r=1}^{s} \frac{\partial f}{\partial F_r} \frac{\partial F_r}{\partial x_j} dt = \sum_{r=1}^{s} \frac{\partial F_r}{\partial x_j} \int_{-\infty}^{x_i} \frac{\partial f}{\partial F_r} dt \quad \text{for} \quad i \neq j,$$

and

$$\frac{\partial y_i}{\partial x_i} = f(x_i \mid F_1, F_2, \ldots, F_s) + \sum_{r=1}^{s} \frac{\partial F_r}{\partial x_i} \int_{-\infty}^{x_i} \frac{\partial f}{\partial F_r} dt \quad \text{for} \quad i = j.$$

In matrix notation we may write this

$$\left[\frac{\partial(y_1, \ldots, y_n)}{\partial(x_1, \ldots, x_n)}\right] = \Delta_I + \left[\frac{\partial F_j}{\partial x_k}\right] \left[\int_{-\infty}^{x_j} \frac{\partial f}{\partial F_k} dt\right].$$

Δ_I is a diagonal matrix with diagonal elements $f(x_i \mid F_1, \ldots, F_s)$. $[\partial F_j / \partial x_k]$ is an $n \times s$ matrix, $\partial F_j / \partial x_k$ being the element in the kth row and the jth column. $\left[\int_{-\infty}^{x_j} \frac{\partial f}{\partial F_k} dt\right]$ is an $s \times n$ matrix, $\int_{-\infty}^{x_j} \frac{\partial f}{\partial F_k} dt$ being the element in the kth row and the jth column. The rank of the matrix $\left[\frac{\partial(y_1, \ldots, y_n)}{\partial(x_1, \ldots, x_n)}\right]$ is not immediately obvious. In general it may be noted that it will be at least $n - s$, and that it will be less than n. A study of particular cases leads us to believe that where the s sample estimates are algebraic functions each of the other, as for example, the sample moment coefficients, then there will be s independent relationships between the y's, and the rank of the matrix will be $n - s$. Where the sample estimates are not functions of one another, as for example in the case of the median and the standard deviation, the matrix rank will be between $n - s$ and n. We have not been able to prove this in general but it should not be impossible to do so.

We shall assume that there are s independent relationships between the variables $y_1, y_2, \ldots, y_n$. Under this last assumption we have

$$\frac{\partial(y_1, \ldots, y_{n-s}, F_1, F_2, \ldots, F_s)}{\partial(x_1, \ldots, x_n)} = \frac{\partial(F_1, \ldots, F_s)}{\partial(x_{n-s+1}, \ldots, x_n)} \prod_{i=1}^{n-s} f(x_i \mid F_1, F_2, \ldots, F_s)$$

(provided partial differentiation is, in fact, possible), whence, since

$$p(x_1, x_2, \ldots, x_n \mid \theta_1, \theta_2, \ldots, \theta_s) = \prod_{i=1}^{n} f(x_i \mid \theta_1, \theta_2, \ldots, \theta_s),$$

the joint probability law of $y_1, y_2, \ldots, y_{n-s}, F_1 \ldots, F_s$, may be written

$$p(y_1, y_2, \ldots, y_{n-s}, F_1, F_2, \ldots, F_s \mid \theta_1, \theta_2, \ldots, \theta_s) = \frac{\prod_{i=1}^{n} f(x_i \mid \theta_1, \theta_2, \ldots, \theta_s)}{\prod_{i=1}^{n-s} f(x_i \mid F_1, F_2, \ldots, F_s)} \left| \frac{\partial(F_1, \ldots, F_s)}{\partial(x_{n-s+1}, \ldots, x_n)} \right|^{-1}.$$

Alternative expressions for the joint-probability law may be obtained by using the relationship $\prod_{i=1}^{n-s} f(x_i \mid \theta_1, \ldots, \theta_s) = p(x_1, \ldots, x_{n-s} \mid \theta_1, \ldots, \theta_s)$.

Substituting in the right-hand side of the joint law we have

$$\begin{aligned} &p(y_1, y_2, \ldots, y_{n-s}, F_1, F_2, \ldots, F_s \mid \theta_1, \theta_2, \ldots, \theta_s) \\ &\quad = \prod_{i=1}^{n-s} \frac{f(x_i \mid \theta_1, \theta_2, \ldots, \theta_s)}{f(x_i \mid F_1, F_2, \ldots, F_s)} \prod_{i=n-s+1}^{n} f(x_i \mid \theta_1, \theta_2, \ldots, \theta_s) \left| \frac{\partial(F_1, \ldots, F_s)}{\partial(x_{n-s+1}, \ldots, x_n)} \right|^{-1} \\ &\quad = \prod_{i=1}^{n-s} \frac{f(x_i \mid \theta_1, \theta_2, \ldots, \theta_s)}{f(x_i \mid F_1, F_2, \ldots, F_s)} p(F_1, F_2, \ldots, F_s \mid x_1, \ldots, x_{n-s+1}, \theta_1, \ldots, \theta_s) \\ &\quad = \frac{p(x_1, \ldots, x_{n-s}, F_1, F_2, \ldots, F_s \mid \theta_1, \ldots, \theta_s)}{\prod_{i=1}^{n-s} f(x_i \mid F_1, F_2, \ldots, F_s)}. \end{aligned}$$

4. In the previous section formal solutions only of the problem have been set down. For any particular case the analysis becomes somewhat complicated. Accordingly, in order to obtain a clear idea of the kinds of distributions arising, we shall first confine ourselves to the discussion of (α) the special case where only location and scale parameters appear in the probability law of the x's and (β) the distribution of single y_i. Under (β) we may note that

$$y_i = \int_{-\infty}^{x_i} f(t \mid F_1, F_2, \ldots, F_s)\, dt = Z_i(x_i, F_1, \ldots, F_s), \quad \text{say.}$$

If the distribution of Z_i is known, the distribution of y_i may be found immediately, and in particular if we can write

$$y_i = Z_i(x_i, F_1, \ldots, F_s) = \int_{-\infty}^{z_i(x_i, F_1, \ldots, F_s)} g(t)\, dt,$$

then $$p(y_i) = (g(z_i))^{-1} p(z_i).$$

5. It is not uncommon in statistical practice to find probability laws which are completely specified by a single parameter for location and a single parameter for scale. The normal curve is, of course, the classic example. Let ξ be the parameter of location, σ the scale parameter, and write

$$p(x) = f(x \mid \xi, \sigma).$$

If $$y_i = \int_{-\infty}^{x_i} f(t \mid \xi, \sigma)\, dt = \int_{-\infty}^{(x_i - \xi)/\sigma} f(t \mid 0, 1)\, dt = \int_{-\infty}^{(x_i - \xi)/\sigma} f(t)\, dt, \quad \text{say,}$$

then we may write $$\frac{x_i - \xi}{\sigma} = \phi(y_i).$$

Either ξ, or σ, or both may be estimated from the observed values of the random variables x. We treat two distinct cases.

Case (i): σ known and ξ estimated

We first suppose that the scaling parameter is known but that it is necessary to estimate a central measure of location. Suppose this to be a function $M(x_1, x_2, \ldots, x_n)$ which may be written for brevity $M(x)$. We have

$$y_i = \int_{-\infty}^{x_i} f(t \mid M(x), \sigma)\, dt = \int_{-\infty}^{[x_i - M(x)]/\sigma} f(t)\, dt. \tag{1}$$

so
$$\frac{x_i - M(x)}{\sigma} = \phi(y_i).$$

It follows that
$$M[\phi(y)] = 0,$$

provided M satisfies the usual conditions for a measure of location.* In this case therefore there is one relation between the y_i's and $\left[\frac{\partial(y_1, \ldots, y_n)}{\partial(x_1, \ldots, x_n)}\right]$ is of rank $n-1$. Using the general formula obtained in § 3, the joint-probability law of $y_1, y_2, \ldots, y_{n-1}, \bar{x}$, is

$$p(y_1, y_2, \ldots, y_{n-1}, \bar{x}) = n \frac{\prod_{i=1}^{n} f(x_i \mid \xi, \sigma)}{\prod_{i=1}^{n-1} f(x_i \mid \bar{x}, \sigma)}.$$

The distribution of any individual y_i is simply obtained. For, since

$$y_i = \int_{-\infty}^{[x_i - M(x)]/\sigma} f(t)\, dt,$$

we have
$$p(y_i) = \left[f\left(\frac{x_i - M(x)}{\sigma}\right)\right]^{-1} p\left(\frac{x_i - M(x)}{\sigma}\right)$$

and if the distribution of $x_i - M(x)$ is known, the distribution of y_i follows immediately.

Case (ii): both ξ and σ estimated

Assume that ξ is estimated as before by $M(x_1, x_2, \ldots, x_n) = M(x)$. Since now σ is also unknown, suppose that it is estimated from the sample values by a measure of dispersion, say
$$D(x_1, x_2, \ldots, x_n) = D(x).$$

We have
$$y_i = \int_{-\infty}^{x_i} f(t \mid M(x), D(x))\, dt,$$

and
$$\frac{x_i - M(x)}{D(x)} = \phi(y_i).$$

Provided $D(x)$ is a function of the quantities $x_i - M(x)$ and satisfies the usual conditions for a measure of dispersion,† the y_i's must now satisfy the two conditions,
$$M[\phi(y)] = 0; \quad D[\phi(y)] = 1.$$

The matrix $\left[\frac{\partial(y_1, \ldots, y_n)}{\partial(x_1, \ldots, x_n)}\right]$ is then of rank $n-2$. The conditions are satisfied, for example, if $M(x) = \bar{x}$ and $D(x) = \left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{1}{2}}$. By an argument precisely similar to that of case (i) we have that
$$p(y_i) = \left[f\left(\frac{x_i - M(x)}{D(x)}\right)\right]^{-1} p\left(\frac{x_i - M(x)}{D(x)}\right),$$

* $M(x_1 + a, x_2 + a, \ldots, x_n + a) = M(x_1, x_2, \ldots, x_n) + a$; $M(x, x, x, \ldots, x) \equiv x$.

† (i) $D(x_1 + a, x_2 + a, \ldots, x_n + a) = D(x_1, x_2, \ldots, x_n)$;
(ii) $D(x, x, \ldots, x) = 0$;
(iii) $D(kx_1, kx_2, \ldots, kx_n) = |k|\, D(x_1, x_2, \ldots, x_n)$.

where
$$y_i = \int_{-\infty}^{[x_i - M(x)]/D(x)} f(t)\,dt. \tag{2}$$

It is seen that y_i is rectangularly distributed, i.e. $p(y_i) = 1$, if and only if

$$f\left(\frac{x_i - M(x)}{D(x)}\right) = p\left(\frac{x_i - M(x)}{D(x)}\right).$$

This condition is not likely to be satisfied.

For both cases (i) and (ii) it may be noted that $p(y_i)$ is the ratio of two probability laws with a transformation of variables given by (1) and (2) respectively, and that *neither of these two probability laws depends on ξ or on σ.*

6. *Example I.* Let

$$p(x) = \frac{1}{\sqrt{(2\pi)}\,\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\xi}{\sigma}\right)^2\right] = f(x \mid \xi, \sigma).$$

and as in the previous section consider two cases.

For case (i) there are many good statistical reasons for choosing

$$M(x) = \bar{x}$$

for the estimate of ξ for this probability law. In the notation of § 4

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

and
$$p\left(\frac{x_i - \bar{x}}{\sigma}\right) = p(z_i) = \frac{1}{\sqrt{(2\pi)}}\sqrt{\frac{n}{n-1}}\exp\left[-\frac{nz_i^2}{2(n-1)}\right].$$

Applying the results of the preceding section we shall have

$$p(y_i) = \left(\frac{1}{\sqrt{(2\pi)}}\exp\left[-\frac{z_i^2}{2}\right]\right)^{-1}\left(\frac{1}{\sqrt{(2\pi)}}\sqrt{\frac{n}{n-1}}\exp\left[-\frac{nz_i^2}{2(n-1)}\right]\right) = \sqrt{\frac{n}{n-1}}\exp\left[-\frac{z_i^2}{2(n-1)}\right], \tag{3}$$

where $z_i = \phi(y_i)$ is defined by
$$y_i = \frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{z_i} e^{-\frac{1}{2}t^2}\,dt.$$

Clearly $p(y_i)$ has a maximum value $\sqrt{n}/(n-1)$ at $z_i = 0$, i.e. when $y = \frac{1}{2}$, and the probability law is symmetrical about this point. $p(y_i)$ is zero at the points $y_i = 0 (z_i = -\infty)$ and $y_i = 1 (z_i = +\infty)$. A graph of the function for three different values of n is given in Fig. 1. In order to compare $p(y_i)$ with the rectangular distribution we may find the points at which the curve crosses it. This will be when $p(y_i) = 1$ or when

$$\frac{z_i^2}{2(n-1)} = -\tfrac{1}{2}\log\left(1 - \frac{1}{n}\right).$$

Expanding the logarithm as a series we have that

$$z_i^2 \doteq 1 - \frac{1}{2n} + \frac{1}{12n^2} - \ldots,$$

or, for n moderately large, z_i is nearly equal to ± 1. It follows that $p(y_i) = 1$ when $y_i \doteq 0{\cdot}159$ or $0{\cdot}841$.

In case (ii) for the same $p(x)$, assume

$$M(x) = \bar{x}, \quad D(x) = s = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{\frac{1}{2}}.$$

As before, write $$z_i = \frac{x_i - \bar{x}}{s},$$

then $$p(z_i) = \frac{\sqrt{n}}{(n-1)\,B\left(\frac{1}{2}, \frac{n-2}{2}\right)} \left(1 - \frac{n z_i^2}{(n-1)^2}\right)^{\frac{1}{2}(n-4)} \quad \text{for} \quad -\frac{n-1}{\sqrt{n}} < z_i < +\frac{n-1}{\sqrt{n}}.$$

It follows that

$$p(y_i) = \left(\frac{1}{\sqrt{(2\pi)}} e^{-\frac{1}{2}z_i^2}\right)^{-1} \left(\frac{\sqrt{n}}{(n-1)\,B\left(\frac{1}{2}, \frac{n-2}{2}\right)} \left(1 - \frac{n z_i^2}{(n-1)^2}\right)^{\frac{1}{2}(n-4)}\right)$$

$$= \frac{\sqrt{(2\pi n)}}{n-1} \frac{1}{B(\frac{1}{2}, \frac{1}{2}(n-2))} \left(1 - \frac{n z_i^2}{(n-1)^2}\right)^{\frac{1}{2}(n-4)} e^{\frac{1}{2}z_i^2}, \tag{4}$$

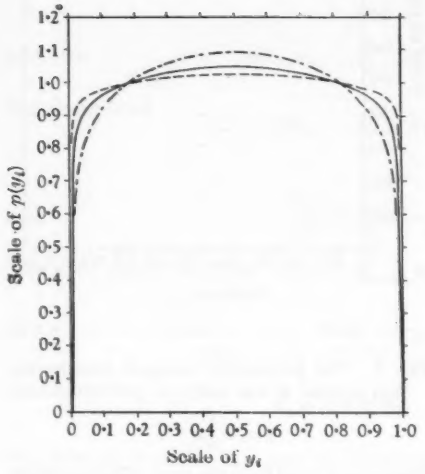


Fig. 1

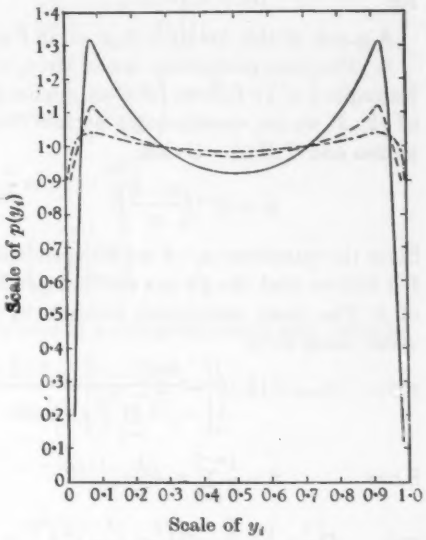


Fig. 2

Fig. 1. The probability integral transformation applied to the normal curve with estimated mean.
Fig. 2. The probability integral transformation applied to the normal curve with estimated mean and standard deviation.

$$\left(\text{Maxima approximately at } y_i = \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{\pm\sqrt{[2+(1/n)]}} e^{-\frac{1}{2}t^2}\,dt.\right)$$

where, as before, $$y_i = \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{z_i} e^{-\frac{1}{2}t^2}\,dt.$$

A graph of this function, for the same sample sizes considered in case (i), is given in Fig. 2.

7. *Example II.* Let x be distributed as χ^2 with two degrees of freedom, i.e. let

$$p(x) = \frac{1}{\theta} e^{-x/\theta} = f(x \mid \theta) \quad \text{for} \quad x > 0.$$

If we estimate θ by $M(x) = \bar{x}$,

then $$y_i = \frac{1}{\bar{x}}\int_0^{x_i} e^{-t/\bar{x}}\,dt = 1 - e^{-x_i/\bar{x}}.$$

In this case we know, writing $u_i = x_i/\bar{x}$,

that $$p(u_i) = \frac{n-1}{n}\left(1 - \frac{u_i}{n}\right)^{n-2} \quad (0 < u_i < n).$$

Following the procedure of the previous sections, we have that

$$p(y_i) = \frac{n-1}{n}\frac{1}{1-y_i}\left[1 + \frac{1}{n}\log(1-y_i)\right]^{n-2}$$

for $$0 < y_i < 1 - e^{-n}. \qquad (5)$$

A graph of this function is given in Fig. 3.

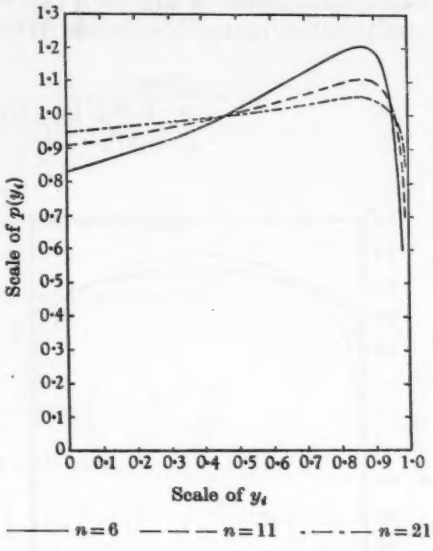


Fig. 3. The probability integral transformation applied to the law

$$p(x) = \frac{1}{\theta}e^{-x/\theta},$$

θ being estimated from the data. (The maximum is at $1-e^{-2} = 0{\cdot}865$ whatever be n.)

8. The joint-probability law of the y_i's for Example I of § 6 follows from an application of § 3. If we are considering a normal distribution and if $M(x) = \bar{x}$ then

$$y_i = \phi^{-1}\left(\frac{x_i - \bar{x}}{\sigma}\right).$$

Since the quantities $x_i - \bar{x}$ are independent of $\bar{x}$ it follows that the y's are also independent of $\bar{x}$. The most convenient formula to use would seem to be

$$p(y_1, \ldots, y_{n-1}, \bar{x} \mid \xi, \sigma) = \frac{p(x_1, \ldots, x_{n-1}, \bar{x} \mid \xi, \sigma)}{\prod_{i=1}^{n-1} f(x_i \mid \bar{x}, \sigma)}.$$

Since $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n-1} x_i + \frac{1}{n}x_n$$

and $$\mathscr{E}\left(\frac{1}{n}x_n\right) = \frac{\xi}{n}, \quad \mathscr{E}\left(\left(\frac{1}{n}x_n\right)^2 - \left(\frac{\xi}{n}\right)^2\right)^{\frac{1}{2}} = \frac{\sigma}{n},$$

it is clear that

$$p(\bar{x} \mid x_1, \ldots, x_{n-1}, \xi, \sigma) = \frac{n}{\sqrt{(2\pi)}\,\sigma}\exp\left\{-\frac{1}{2\sigma^2}\left[(\bar{x} - \xi) - \sum_{i=1}^{n-1}(x_i - \bar{x})\right]^2\right\}.$$

Hence

$$p(x_1, \ldots, x_{n-1}, \bar{x} \mid \xi, \sigma) = \frac{n}{(\sqrt{(2\pi)}\,\sigma)^n}\exp\left(-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{n-1}(x_i - \xi)^2 + \left[(\bar{x} - \xi) - \sum_{i=1}^{n-1}(x_i - \bar{x})\right]^2\right\}\right),$$

and $$\prod_{i=1}^{n-1} f(x_i \mid \bar{x}, \sigma) = \frac{1}{(\sqrt{(2\pi)}\,\sigma)^{n-1}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n-1}(x_i - \bar{x})^2\right].$$

The joint-probability law of $y_1, y_2, \ldots, y_{n-1}$ and $\bar{x}$ will be

$$p(y_1, \ldots, y_{n-1}, \bar{x} \mid \xi, \sigma) = \frac{n}{\sqrt{(2\pi)}\,\sigma}\exp\left(-\frac{1}{2\sigma^2}\left\{n(\bar{x} - \xi)^2 + \left[\sum_{i=1}^{n-1}(x_i - \bar{x})\right]^2\right\}\right)$$

$$= \frac{n}{\sqrt{(2\pi)}\,\sigma}\exp\left[-\frac{n(\bar{x} - \xi)^2}{2\sigma^2}\right]\exp\left[-\frac{1}{2}\left\{\sum_{i=1}^{n-1}\phi(y_i)\right\}^2\right],$$

$\phi(y_i)$ being defined as in (1). Integrate out for $\bar{x}$, and we have

$$p(y_1, \ldots, y_{n-1} \mid \xi, \sigma) = \sqrt{n} \exp\left[-\frac{1}{2}\left\{\sum_{i=1}^{n-1} \phi(y_i)\right\}^2\right]. \tag{6}$$

It will be noted that this joint probability law is independent of both ξ and σ.

9. The exponential law discussed in § 7 differs from the examples in which the normal law was used in that in this particular example a measure of location is used to estimate a scale parameter. We have

$$p(x) = \frac{1}{\theta} e^{-x/\theta},$$

whence, since

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n-1} x_i + \frac{1}{n} x_n,$$

it is seen that

$$p(\bar{x} \mid x_1, \ldots, x_{n-1}, \theta) = \frac{n}{\theta} \exp\left(-\frac{n}{\theta}\left[\bar{x} - \frac{1}{n}\sum_{i=1}^{n-1} x_i\right]\right),$$

and also

$$p(x_1, x_2, \ldots, x_{n-1} \mid \theta) = \frac{1}{\theta^{n-1}} \exp\left(-\frac{\sum_{i=1}^{n-1} x_i}{\theta}\right).$$

It follows that

$$p(x_1, \ldots, x_{n-1}, \bar{x} \mid \theta) = \frac{n}{\theta^n} \exp\left(-\frac{n\bar{x}}{\theta}\right),$$

and

$$\prod_{i=1}^{n-1} f(x_i \mid \bar{x}) = \frac{1}{\bar{x}^{n-1}} \exp\left(-\frac{\sum_{i=1}^{n-1} x_i}{\bar{x}}\right).$$

The joint-probability law of $y_1, \ldots, y_{n-1}$, and $\bar{x}$ follows in a straightforward way, namely

$$p(y_1, \ldots, y_{n-1}, \bar{x} \mid \theta) = \frac{n}{\theta}\left(\frac{\bar{x}}{\theta}\right)^{n-1} \exp\left(-\left[\frac{n\bar{x}}{\theta} - \frac{\sum_{i=1}^{n-1} x_i}{\bar{x}}\right]\right).$$

Remembering that

$$y_i = 1 - e^{-x_i/\bar{x}} \quad \text{or} \quad \frac{x_i}{\bar{x}} = -\log(1-y_i)$$

the joint-probability law may be rewritten

$$p(y_1, \ldots, y_{n-1}, \bar{x} \mid \theta) = \frac{n}{\theta}\left(\frac{\bar{x}}{\theta}\right)^{n-1} e^{-n\bar{x}/\theta}\left[\frac{1}{\prod_{i=1}^{n-1}(1-y_i)}\right],$$

where

$$\prod_{i=1}^{n-1}[-\log(1-y_i)] < n \quad \text{or} \quad \prod_{i=1}^{n-1}(1-y_i) < e^{-n}.$$

Integrating out with respect to $\bar{x}$,

$$p(y_1, \ldots, y_{n-1} \mid \theta) = \frac{(n-1)!}{(n-1)^{n-1}} \frac{1}{\prod_{i=1}^{n-1}(1-y_i)}, \quad \prod_{i=1}^{n-1}(1-y_i) < e^{-n}, \tag{7}$$

again a result which is independent of the parameter of the probability law.

10. The results of this investigation, which we have carried out partly in the general and partly in the particular, are obviously incomplete and should be succeeded by a fuller inquiry which would clear up the doubtful points which we have had to pass over, and possibly extend the general theory a little further. We feel that none of the questions which

have been raised in the course of this inquiry are insoluble by algebraic analysis but it is uncertain whether it is profitable to proceed with the fuller inquiry until some of the statistical implications of what has been done become more clear. For example, we have noted that given n independent random variables, x, if s sample moments are calculated from them and used as estimates of the parameters of the probability law, then it appears that there will be s independent relationships between the y's. Thus in this case the point $y_1, y_2, \ldots, y_n$ is constrained to move in an $n-s$ dimensioned space within an n dimensioned cube, and we have the exact analogue to the loss of degrees of freedom with χ^2 when the parameters have to be estimated from the data. What is not clear is how the y's are constrained when the sample estimates of the parameters are not the sample moments, and while this situation may not often be met with in practice, yet it should be explored.

When the parameters of location and scale are estimated from the data it is clear that the distribution of any individual y_i, and the joint-probability law of the y's also, will not be dependent on these unknown parameters of the probability law of the x's, but will depend on the functional form of that law. This result appears capable of extension for the case when higher sample moments are also used for estimating parameters. This being so, there would seem to be two ways in which the joint probability law of the y's may be utilized in statistical applications. First, it should be possible mathematically to form certain broad classes of functions for each of which the joint-probability laws of the y's would be approximately the same, or second one may seek for some transformation of variables so that instead of the correlated y_i we obtain $n-s$ new independent variables following some distributions which are independent of the original $p(x)$. Both these methods of attack may lead to results which will only be valid for large samples, but provided the results in either case have sufficient algebraic simplicity they should make possible certain generalizations in statistical analysis of which Neyman's 'smooth' test for goodness of fit is only one important example.

REFERENCES

FISHER, R. A. (1932). *Statistical Methods for Research Workers*, § 21·1.
NEYMAN, J. (1937). *Skand. AktuarTidskr.* **20**, 149.
PEARSON, E. S. (1938). *Biometrika*, **30**, 134.
PEARSON, K. (1933). *Biometrika*, **25**, 379.

A TABLE FOR THE CALCULATION OF WORKING PROBITS AND WEIGHTS IN PROBIT ANALYSIS

By D. J. FINNEY (*Lecturer in the Design and Analysis of Scientific Experiment, University of Oxford*) and W. L. STEVENS (*Admiralty*)

The estimation of the parameters of a distribution of individual tolerances, from data relating to numbers of subjects manifesting a characteristic quantal response at different levels of a stimulus, is a problem frequently encountered in the application of statistical science to dose-mortality studies, biological assay, detonation of explosives, and other problems. A typical situation is that of exposing batches of insects to various doses of an insecticide, recording the proportion killed at each level of dose and then requiring to estimate the mean tolerance (or median lethal dose) of individual insects and the variance of the tolerance distribution. Gaddum (1933) and Bliss (1935*a*, *b*; 1938) have been instrumental in developing a method, that of the *probit transformation*, which greatly simplifies the calculations necessary to the estimation. The exact statistical analysis appropriate to the transformation was first shown by Fisher (1935), and the theory and uses of the method have been discussed fully in many subsequent publications (Finney, 1947*a*, *b*).

Tables required in the practice of the method, in sufficient detail for most purposes, have been given by various writers (Fisher & Yates, 1943; Finney, 1947*a*). Occasionally, however, the statistician needs values of the various functions at finer intervals of the argument, and for his benefit the following Table has been prepared. A brief account of the tabulated functions will suffice for all who are familiar with the probit method; those who require fuller information on the theory and analysis should consult the list of References.

Given a proportion P, and its complement $Q = 1 - P$, the probit of P is, to all intents and purposes, the deviate from the mean which divides the normal curve of unit variance in the ratio $P:Q$. In the formal definition, however, 5 is added to the deviate in order to avoid the necessity of computing with negative numbers. The advantage of this modification may be questioned, but it is now well established and will be adopted here. The probit, Y, of the proportion P is thus defined by

$$P = \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{Y-5} e^{-\frac{1}{2}u^2} du.$$

The standard method of analysis makes use of the *maximum and minimum working probits*,

$$Y_{\max.} = Y + \frac{Q}{Z}$$

and

$$Y_{\min.} = Y - \frac{P}{Z};$$

and also of the *range*,

$$1/Z,$$

where

$$Z = \frac{1}{\sqrt{(2\pi)}} e^{-\frac{1}{2}(Y-5)^2}.$$

If n subjects receive the same stimulus, and r of them show the characteristic response, the empirical value for the proportion responding is

$$p = r/n;$$

the complement of this is denoted by $q = 1-p$. The probits of a set of values of p should be approximately linearly related to x, the measure of the stimulus, and a line fitted by eye may be used to give a corresponding set of *expected probits*, Y. The *working probit* corresponding to each proportion is next calculated, from either

$$y = Y + Q/Z - q/Z,$$

or

$$y = Y - P/Z + p/Z,$$

using tabulated values of the maximum or the minimum working probit (whichever is the more convenient) and the range. An improved set of expected probits is then derived from the weighted linear regression equation of working probits on x, each y being assigned a weight, nw, where the *weighting coefficient*, w, is defined as

$$w = Z^2/PQ.$$

The process may be repeated with the new set of Y values. The iteration converges to give a linear regression equation which is an estimate of

$$Y = 5 + (x-\mu)/\sigma,$$

where μ is the mean and σ the standard deviation of the tolerance distribution. The method depends upon an assumption that the stimulus is measured on a scale for which individual tolerances are normally distributed: often the logarithm of 'dose' rather than dose itself is taken as x, in order to satisfy this condition more closely.

The Table which follows gives $Y_{\text{max.}}$ for $Y = 3{\cdot}58(0{\cdot}01)\ 9{\cdot}00$, $Y_{\text{min.}}$ for $Y = 1{\cdot}00(0{\cdot}01)\ 6{\cdot}42$, $1/Z$ and w for $Y = 1{\cdot}00(0{\cdot}01)\ 9{\cdot}00$, all to four places of decimals. Below $Y = 3{\cdot}58$, $Y_{\text{max.}}$ exceeds 10·00, and above $Y = 6{\cdot}52$, $Y_{\text{min.}}$ is negative; it is then almost always more convenient to calculate working probits from the other function, but the function not tabulated can easily be obtained from the relationship

$$Y_{\text{max.}} - Y_{\text{min.}} = 1/Z.$$

Between 3·58 and 6·42 both functions are tabulated. In order to save space the Table is arranged in parallel forward- and backward-reading columns; for the arguments Y and $(10-Y)$ values of $1/Z$ and w are the same, and simple relations exist between $Y_{\text{max.}}$ and $Y_{\text{min.}}$. All entries have been calculated to six or more places of decimals and rounded to four, except that w between $Y = 2{\cdot}7$ and $Y = 7{\cdot}3$ was obtained by collating two existing tables and checking discrepancies.

The values of P and Z, from which the present Table has been calculated, were taken from *Tables of the Probability Function*, Vol. II (1942), published by the Federal Works Project Administration for the City of New York. Values of Q/Z have been taken from, or checked against, W. F. Sheppard's table, published as *The Probability Integral* (1939), Vol. VII of the *British Association Mathematical Tables*.

Example

In a batch of 281 insects receiving the same dose of insecticide, 119 are killed. The provisional probit regression line gives an expected probit of 4·61 for this dose; find the working probit and the weight to be attached to the observation.

Expected probit Y	Maximum working probit $Y+Q/Z$	Minimum working probit $Y-P/Z$	Range $1/Z$	Weighting coefficient Z^2/PQ			
5·00	6·2533	3·7467	2·5066	0·6366	6·2533	3·7467	5·00
·01	·2534	·7466	·5068	·6366	·2534	·7466	4·99
·02	·2536	·7465	·5071	·6365	·2535	·7464	·98
·03	·2539	·7461	·5078	·6364	·2539	·7461	·97
·04	·2543	·7457	·5086	·6362	·2543	·7457	·96
5·05	6·2548	3·7450	2·5098	0·6360	6·2550	3·7452	4·95
·06	·2555	·7444	·5111	·6358	·2556	·7445	·94
·07	·2563	·7435	·5128	·6355	·2565	·7437	·93
·08	·2572	·7425	·5147	·6351	·2575	·7428	·92
·09	·2582	·7414	·5168	·6347	·2586	·7418	·91
5·10	6·2593	3·7401	2·5192	0·6343	6·2599	3·7407	4·90
·11	·2605	·7387	·5218	·6338	·2613	·7395	·89
·12	·2618	·7371	·5247	·6333	·2629	·7382	·88
·13	·2632	·7353	·5279	·6327	·2647	·7368	·87
·14	·2647	·7334	·5313	·6321	·2666	·7353	·86
5·15	6·2664	3·7314	2·5350	0·6314	6·2686	3·7336	4·85
·16	·2681	·7292	·5389	·6307	·2708	·7319	·84
·17	·2699	·7268	·5431	·6300	·2732	·7301	·83
·18	·2718	·7242	·5476	·6292	·2758	·7282	·82
·19	·2738	·7215	·5523	·6283	·2785	·7262	·81
5·20	6·2759	3·7186	2·5573	0·6274	6·2814	3·7241	4·80
·21	·2781	·7156	·5625	·6265	·2844	·7219	·79
·22	·2804	·7124	·5680	·6255	·2876	·7196	·78
·23	·2828	·7090	·5738	·6245	·2910	·7172	·77
·24	·2853	·7054	·5799	·6234	·2946	·7147	·76
5·25	6·2878	3·7016	2·5862	0·6223	6·2984	3·7122	4·75
·26	·2905	·6977	·5928	·6211	·3023	·7095	·74
·27	·2932	·6935	·5997	·6199	·3065	·7068	·73
·28	·2960	·6892	·6068	·6187	·3108	·7040	·72
·29	·2989	·6846	·6143	·6174	·3154	·7011	·71
5·30	6·3018	3·6798	2·6220	0·6161	6·3202	3·6982	4·70
·31	·3049	·6749	·6300	·6147	·3251	·6951	·69
·32	·3080	·6697	·6383	·6133	·3303	·6920	·68
·33	·3112	·6643	·6469	·6119	·3357	·6888	·67
·34	·3145	·6587	·6558	·6104	·3413	·6855	·66
5·35	6·3178	3·6528	2·6650	0·6088	6·3472	3·6822	4·65
·36	·3213	·6469	·6744	·6072	·3531	·6787	·64
·37	·3248	·6406	·6842	·6056	·3594	·6752	·63
·38	·3283	·6340	·6943	·6040	·3660	·6717	·62
·39	·3320	·6273	·7047	·6023	·3727	·6680	·61
5·40	6·3357	3·6203	2·7154	0·6005	6·3797	3·6643	4·60
·41	·3394	·6130	·7264	·5987	·3870	·6606	·59
·42	·3433	·6055	·7378	·5969	·3945	·6567	·58
·43	·3472	·5978	·7494	·5951	·4022	·6528	·57
·44	·3512	·5898	·7614	·5932	·4102	·6488	·56
5·45	6·3552	3·5815	2·7737	0·5912	6·4185	3·6448	4·55
·46	·3593	·5729	·7864	·5893	·4271	·6407	·54
·47	·3635	·5641	·7994	·5872	·4359	·6365	·53
·48	·3677	·5550	·8127	·5852	·4450	·6323	·52
·49	·3720	·5456	·8264	·5831	·4544	·6280	·51
5·50	6·3764	3·5360	2·8404	0·5810	6·4640	3·6236	4·50

$1/Z$ Range	Z^2/PQ Weighting coefficient	$Y+Q/Z$ Maximum working probit	$Y-P/Z$ Minimum working probit	Y Expected probit

Expected probit Y	Maximum working probit $Y+Q/Z$	Minimum working probit $Y-P/Z$	Range $1/Z$	Weighting coefficient Z^2/PQ			
5·50	6·3764	3·5360	2·8404	0·5810	6·4640	3·6236	4·50
·51	·3808	·5260	·8548	·5788	·4740	·6192	·49
·52	·3852	·5157	·8695	·5766	·4843	·6148	·48
·53	·3898	·5052	·8846	·5744	·4948	·6102	·47
·54	·3944	·4943	·9001	·5722	·5057	·6056	·46
5·55	6·3990	3·4831	2·9159	0·5699	6·5169	3·6010	4·45
·56	·4037	·4715	·9322	·5675	·5285	·5963	·44
·57	·4085	·4597	·9488	·5652	·5403	·5915	·43
·58	·4133	·4475	·9658	·5628	·5525	·5867	·42
·59	·4181	·4349	·9832	·5603	·5651	·5819	·41
5·60	6·4230	3·4220	3·0010	0·5579	6·5780	3·5770	4·40
·61	·4280	·4088	·0192	·5554	·5912	·5720	·39
·62	·4330	·3952	·0378	·5529	·6048	·5670	·38
·63	·4381	·3812	·0569	·5503	·6188	·5619	·37
·64	·4432	·3669	·0763	·5477	·6331	·5568	·36
5·65	6·4484	3·3522	3·0962	0·5451	6·6478	3·5516	4·35
·66	·4536	·3370	·1166	·5425	·6630	·5464	·34
·67	·4588	·3214	·1374	·5398	·6786	·5412	·33
·68	·4641	·3055	·1586	·5371	·6945	·5359	·32
·69	·4695	·2892	·1803	·5343	·7108	·5305	·31
5·70	6·4749	3·2724	3·2025	0·5316	6·7276	3·5251	4·30
·71	·4803	·2551	·2252	·5288	·7449	·5197	·29
·72	·4858	·2375	·2483	·5260	·7625	·5142	·28
·73	·4914	·2194	·2720	·5232	·7806	·5086	·27
·74	·4969	·2008	·2961	·5203	·7992	·5031	·26
5·75	6·5026	3·1819	3·3207	0·5174	6·8181	3·4974	4·25
·76	·5082	·1623	·3459	·5145	·8377	·4918	·24
·77	·5139	·1423	·3716	·5116	·8577	·4861	·23
·78	·5197	·1219	·3978	·5086	·8781	·4803	·22
·79	·5255	·1009	·4246	·5056	·8991	·4745	·21
5·80	6·5313	3·0794	3·4519	0·5026	6·9206	3·4687	4·20
·81	·5372	·0574	·4798	·4996	·9426	·4628	·19
·82	·5431	·0348	·5083	·4965	·9652	·4569	·18
·83	·5490	·0116	·5374	·4935	·9884	·4510	·17
·84	·5550	2·9880	·5670	·4904	7·0120	·4450	·16
5·85	6·5611	2·9638	3·5973	0·4873	7·0362	3·4389	4·15
·86	·5671	·9389	·6282	·4841	·0611	·4329	·14
·87	·5732	·9135	·6597	·4810	·0865	·4268	·13
·88	·5794	·8875	·6919	·4778	·1125	·4206	·12
·89	·5855	·8608	·7247	·4746	·1392	·4145	·11
5·90	6·5917	2·8335	3·7582	0·4714	7·1665	3·4083	4·10
·91	·5980	·8056	·7924	·4682	·1944	·4020	·09
·92	·6043	·7771	·8272	·4650	·2229	·3957	·08
·93	·6106	·7478	·8628	·4617	·2522	·3894	·07
·94	·6169	·7178	·8991	·4585	·2822	·3831	·06
5·95	6·6233	2·6872	3·9361	0·4552	7·3128	3·3767	4·05
·96	·6297	·6558	·9739	·4519	·3442	·3703	·04
·97	·6362	·6238	4·0124	·4486	·3762	·3638	·03
·98	·6426	·5909	·0517	·4453	·4091	·3574	·02
·99	·6491	·5573	·0918	·4420	·4427	·3509	·01
6·00	6·6557	2·5230	4·1327	0·4386	7·4770	3·3443	4·00
			$1/Z$ Range	Z^2/PQ Weighting coefficient	$Y+Q/Z$ Maximum working probit	$Y-P/Z$ Minimum working probit	Y Expected probit

Expected probit Y	Maximum working probit $Y+Q/Z$	Minimum working probit $Y-P/Z$	Range $1/Z$	Weighting coefficient Z^2/PQ			
6·00	6·6557	2·5230	4·1327	0·4386	7·4770	3·3443	4·00
·01	·6623	·4878	·1745	·4353	·5122	·3377	3·99
·02	·6689	·4518	·2171	·4319	·5482	·3311	·98
·03	·6755	·4150	·2605	·4285	·5850	·3245	·97
·04	·6822	·3774	·3048	·4252	·6226	·3178	·96
6·05	6·6888	2·3387	4·3501	0·4218	7·6613	3·3112	3·95
·06	·6956	·2994	·3962	·4184	·7006	·3044	·94
·07	·7023	·2590	·4433	·4150	·7410	·2977	·93
·08	·7091	·2178	·4913	·4116	·7822	·2909	·92
·09	·7159	·1756	·5403	·4082	·8244	·2841	·91
6·10	6·7227	2·1324	4·5903	0·4047	7·8676	3·2773	3·90
·11	·7296	·0883	·6413	·4013	·9117	·2704	·89
·12	·7365	·0432	·6933	·3979	·9568	·2635	·88
·13	·7434	1·9970	·7464	·3944	8·0030	·2566	·87
·14	·7504	·9498	·8006	·3910	·0502	·2496	·86
6·15	6·7573	1·9014	4·8559	0·3876	8·0986	3·2427	3·85
·16	·7643	·8520	·9123	·3841	·1480	·2357	·84
·17	·7714	·8016	·9698	·3807	·1984	·2286	·83
·18	·7784	·7498	5·0286	·3772	·2502	·2216	·82
·19	·7855	·6970	·0885	·3738	·3030	·2145	·81
6·20	6·7926	1·6429	5·1497	0·3703	8·3571	3·2074	3·80
·21	·7997	·5876	·2121	·3669	·4124	·2003	·79
·22	·8068	·5310	·2758	·3634	·4690	·1932	·78
·23	·8140	·4731	·3409	·3600	·5269	·1860	·77
·24	·8212	·4140	·4072	·3565	·5860	·1788	·76
6·25	6·8284	1·3534	5·4750	0·3531	8·6466	3·1716	3·75
·26	·8357	·2916	·5441	·3496	·7084	·1643	·74
·27	·8429	·2282	·6147	·3462	·7718	·1571	·73
·28	·8502	·1635	·6867	·3428	·8365	·1498	·72
·29	·8575	·0972	·7603	·3393	·9028	·1425	·71
6·30	6·8649	1·0295	5·8354	0·3359	8·9705	3·1351	3·70
·31	·8722	0·9602	·9120	·3325	9·0398	·1278	·69
·32	·8796	·8893	·9903	·3291	·1107	·1204	·68
·33	·8870	·8168	6·0702	·3256	·1832	·1130	·67
·34	·8944	·7426	·1518	·3222	·2574	·1056	·66
6·35	6·9019	0·6668	6·2351	0·3188	9·3332	3·0981	3·65
·36	·9093	·5892	·3201	·3155	·4108	·0907	·64
·37	·9168	·5098	·4070	·3121	·4902	·0832	·63
·38	·9243	·4286	·4957	·3087	·5714	·0757	·62
·39	·9318	·3455	·5863	·3053	·6545	·0682	·61
6·40	6·9394	0·2606	6·6788	0·3020	9·7394	3·0606	3·60
·41	·9469	·1736	·7733	·2986	·8264	·0531	·59
·42	·9545	·0847	·8698	·2953	·9153	·0455	·58
·43	·9621		·9684	·2920		·0379	·57
·44	·9697		7·0691	·2887		·0303	·56
6·45	6·9774		7·1720	0·2854		3·0226	3·55
·46	·9850		·2771	·2821		·0150	·54
·47	6·9927		·3845	·2788		·0073	·53
·48	7·0004		·4943	·2756		2·9996	·52
·49	·0081		·6064	·2723		·9919	·51
6·50	7·0158		7·7210	0·2691		2·9842	3·50
			$1/Z$ Range	Z^2/PQ Weighting coefficient	$Y+Q/Z$ Maximum working probit	$Y-P/Z$ Minimum working probit	Y Expected probit

Expected probit Y	Maximum working probit $Y+Q/Z$	Range $1/Z$	Weighting coefficient Z^2/PQ		
6·50	7·0158	7·7210	0·2691	2·9842	3·50
·51	·0236	·8380	·2658	·9764	·49
·52	·0313	·9577	·2626	·9687	·48
·53	·0391	8·0800	·2594	·9609	·47
·54	·0469	·2050	·2563	·9531	·46
6·55	7·0547	8·3327	0·2531	2·9453	3·45
·56	·0625	·4633	·2500	·9375	·44
·57	·0704	·5968	·2468	·9296	·43
·58	·0783	·7333	·2437	·9217	·42
·59	·0861	·8728	·2406	·9139	·41
6·60	7·0940	9·0154	0·2375	2·9060	3·40
·61	·1020	·1613	·2345	·8980	·39
·62	·1099	·3105	·2314	·8901	·38
·63	·1178	·4630	·2284	·8822	·37
·64	·1258	·6190	·2254	·8742	·36
6·65	7·1338	9·7785	0·2224	2·8662	3·35
·66	·1417	·9417	·2194	·8583	·34
·67	·1498	10·1086	·2165	·8502	·33
·68	·1578	10·2794	·2135	·8422	·32
·69	·1658	10·4540	·2106	·8342	·31
6·70	7·1739	10·6327	0·2077	2·8261	3·30
·71	·1819	10·8156	·2049	·8181	·29
·72	·1900	11·0027	·2020	·8100	·28
·73	·1981	11·1941	·1992	·8019	·27
·74	·2062	11·3900	·1964	·7938	·26
6·75	7·2143	11·5905	0·1936	2·7857	3·25
·76	·2224	11·7957	·1908	·7776	·24
·77	·2306	12·0058	·1881	·7694	·23
·78	·2387	12·2208	·1853	·7613	·22
·79	·2469	12·4409	·1826	·7531	·21
6·80	7·2551	12·6662	0·1799	2·7449	3·20
·81	·2633	12·8969	·1773	·7367	·19
·82	·2715	13·1331	·1746	·7285	·18
·83	·2797	13·3750	·1720	·7203	·17
·84	·2880	13·6227	·1694	·7120	·16
6·85	7·2962	13·8764	0·1669	2·7038	3·15
·86	·3045	14·1362	·1643	·6955	·14
·87	·3128	14·4023	·1618	·6872	·13
·88	·3210	14·6749	·1593	·6790	·12
·89	·3293	14·9541	·1568	·6707	·11
6·90	7·3376	15·2402	0·1544	2·6624	3·10
·91	·3460	15·5333	·1519	·6540	·09
·92	·3543	15·8337	·1495	·6457	·08
·93	·3626	16·1414	·1471	·6374	·07
·94	·3710	16·4568	·1448	·6290	·06
6·95	7·3794	16·7800	0·1424	2·6206	3·05
·96	·3877	17·1113	·1401	·6123	·04
·97	·3961	17·4509	·1378	·6039	·03
·98	·4045	17·7989	·1356	·5955	·02
·99	·4129	18·1558	·1333	·5871	·01
7·00	7·4214	18·5216	0·1311	2·5786	3·00
		$1/Z$ Range	Z^2/PQ Weighting coefficient	$Y-P/Z$ Minimum working probit	Y Expected probit

Expected probit Y	Maximum working probit $Y+Q/Z$	Range $1/Z$	Weighting coefficient Z^2/PQ		
7·00	7·4214	18·5216	0·1311	2·5786	3·00
·01	·4298	18·8967	·1289	·5702	2·99
·02	·4382	19·2814	·1268	·5618	·98
·03	·4467	19·6758	·1246	·5533	·97
·04	·4552	20·0803	·1225	·5448	·96
7·05	7·4636	20·4952	0·1204	2·5364	2·95
·06	·4721	20·9207	·1183	·5279	·94
·07	·4806	21·3572	·1163	·5194	·93
·08	·4891	21·8050	·1142	·5109	·92
·09	·4976	22·2644	·1122	·5024	·91
7·10	7·5062	22·7357	0·1103	2·4938	2·90
·11	·5147	23·2194	·1083	·4853	·89
·12	·5232	23·7157	·1064	·4768	·88
·13	·5318	24·2251	·1045	·4682	·87
·14	·5404	24·7478	·1026	·4596	·86
7·15	7·5489	25·2844	0·1007	2·4511	2·85
·16	·5575	25·8352	·0989	·4425	·84
·17	·5661	26·4006	·0971	·4339	·83
·18	·5747	26·9812	·0953	·4253	·82
·19	·5833	27·5772	·0935	·4167	·81
7·20	7·5919	28·1892	0·0918	2·4081	2·80
·21	·6006	28·8177	·0901	·3994	·79
·22	·6092	29·4631	·0884	·3908	·78
·23	·6178	30·1260	·0867	·3822	·77
·24	·6265	30·8069	·0851	·3735	·76
7·25	7·6351	31·5063	0·0834	2·3649	2·75
·26	·6438	32·2249	·0818	·3562	·74
·27	·6525	32·9631	·0802	·3475	·73
·28	·6612	33·7216	·0787	·3388	·72
·29	·6699	34·5010	·0771	·3301	·71
7·30	7·6786	35·3020	0·0756	2·3214	2·70
·31	·6873	36·1251	·0741	·3127	·69
·32	·6960	36·9712	·0727	·3040	·68
·33	·7047	37·8408	·0712	·2953	·67
·34	·7135	38·7348	·0698	·2865	·66
7·35	7·7222	39·6539	0·0684	2·2778	2·65
·36	·7310	40·5988	·0671	·2690	·64
·37	·7397	41·5704	·0656	·2603	·63
·38	·7485	42·5695	·0643	·2515	·62
·39	·7573	43·5970	·0630	·2427	·61
7·40	7·7661	44·6538	0·0617	2·2339	2·60
·41	·7748	45·7407	·0604	·2252	·59
·42	·7836	46·8588	·0591	·2164	·58
·43	·7924	48·0090	·0579	·2076	·57
·44	·8013	49·1924	·0567	·1987	·56
7·45	7·8101	50·4099	0·0555	2·1899	2·55
·46	·8189	51·6628	·0543	·1811	·54
·47	·8277	52·9521	·0532	·1723	·53
·48	·8366	54·2791	·0520	·1634	·52
·49	·8454	55·6448	·0509	·1546	·51
7·50	7·8543	57·0506	0·0498	2·1457	2·50
		$1/Z$ Range	Z^2/PQ Weighting coefficient	$Y-P/Z$ Minimum working probit	Y Expected probit

Expected probit Y	Maximum working probit $Y+Q/Z$	Range $1/Z$	Weighting coefficient Z^2/PQ		
7·50	7·8543	57·0506	0·0498	2·1457	2·50
·51	·8631	58·4978	·0487	·1369	·49
·52	·8720	59·9876	·0476	·1280	·48
·53	·8809	61·5216	·0466	·1191	·47
·54	·8897	63·1011	·0456	·1103	·46
7·55	7·8986	64·7277	0·0446	2·1014	2·45
·56	·9075	66·4028	·0436	·0925	·44
·57	·9164	68·1280	·0426	·0836	·43
·58	·9253	69·9051	·0416	·0747	·42
·59	·9342	71·7357	·0407	·0658	·41
7·60	7·9432	73·6216	0·0398	2·0568	2·40
·61	·9521	75·5646	·0389	·0479	·39
·62	·9610	77·5667	·0380	·0390	·38
·63	·9700	79·6298	·0371	·0300	·37
·64	·9789	81·7559	·0362	·0211	·36
7·65	7·9879	83·9472	0·0354	2·0121	2·35
·66	·9968	86·2059	·0346	·0032	·34
·67	8·0058	88·5342	·0338	1·9942	·33
·68	·0147	90·9344	·0330	·9853	·32
·69	·0237	93·4091	·0322	·9763	·31
7·70	8·0327	95·9607	0·0314	1·9673	2·30
·71	·0417	98·5918	·0307	·9583	·29
·72	·0507	101·3053	·0300	·9493	·28
·73	·0597	104·1038	·0292	·9403	·27
·74	·0687	106·9903	·0285	·9313	·26
7·75	8·0777	109·9679	0·0278	1·9223	2·25
·76	·0867	113·0396	·0272	·9133	·24
·77	·0957	116·2088	·0265	·9043	·23
·78	·1047	119·4788	·0258	·8953	·22
·79	·1138	122·8530	·0252	·8862	·21
7·80	8·1228	126·3352	0·0246	1·8772	2·20
·81	·1318	129·9290	·0240	·8682	·19
·82	·1409	133·6385	·0234	·8591	·18
·83	·1499	137·4676	·0228	·8501	·17
·84	·1590	141·4206	·0222	·8410	·16
7·85	8·1681	145·5018	0·0217	1·8319	2·15
·86	·1771	149·7158	·0211	·8229	·14
·87	·1862	154·0671	·0206	·8138	·13
·88	·1953	158·5609	·0200	·8047	·12
·89	·2044	163·2020	·0195	·7956	·11
7·90	8·2134	167·9957	0·0190	1·7866	2·10
·91	·2225	172·9476	·0185	·7775	·09
·92	·2316	178·0632	·0181	·7684	·08
·93	·2407	183·3485	·0176	·7593	·07
·94	·2498	188·8095	·0171	·7502	·06
7·95	8·2590	194·4526	0·0167	1·7410	2·05
·96	·2681	200·2844	·0162	·7319	·04
·97	·2772	206·3118	·0158	·7228	·03
·98	·2863	212·5418	·0154	·7137	·02
·99	·2955	218·9818	·0150	·7045	·01
8·00	8·3046	225·6395	0·0146	1·6954	2·00
		$1/Z$ Range	Z^2/PQ Weighting coefficient	$Y-P/Z$ Minimum working probit	Y Expected probit

Expected probit Y	Maximum working probit $Y+Q/Z$	Range $1/Z$	Weighting coefficient Z^2/PQ		
8·00	8·3046	225·6395	0·0146	1·6954	2·00
·01	·3137	232·5229	·0142	·6863	1·99
·02	·3229	239·6402	·0138	·6771	·98
·03	·3320	247·0000	·0134	·6680	·97
·04	·3412	254·6114	·0131	·6588	·96
8·05	8·3503	262·4836	0·0127	1·6497	1·95
·06	·3595	270·6262	·0124	·6405	·94
·07	·3687	279·0493	·0120	·6313	·93
·08	·3778	287·7634	·0117	·6222	·92
·09	·3870	296·7792	·0114	·6130	·91
8·10	8·3962	306·1082	0·0110	1·6038	1·90
·11	·4054	315·7619	·0107	·5946	·89
·12	·4146	325·7527	·0104	·5854	·88
·13	·4238	336·0932	·0101	·5762	·87
·14	·4330	346·7966	·0099	·5670	·86
8·15	8·4422	357·8732	0·0096	1·5578	1·85
·16	·4514	369·3477	·0093	·5486	·84
·17	·4606	381·2245	·0090	·5394	·83
·18	·4698	393·5226	·0088	·5302	·82
·19	·4790	406·2580	·0085	·5210	·81
8·20	8·4882	419·4476	0·0083	1·5118	1·80
·21	·4974	433·1086	·0080	·5026	·79
·22	·5067	447·2593	·0078	·4933	·78
·23	·5159	461·9185	·0076	·4841	·77
·24	·5251	477·1059	·0074	·4749	·76
8·25	8·5344	492·8419	0·0071	1·4656	1·75
·26	·5436	509·1479	·0069	·4564	·74
·27	·5529	526·0459	·0067	·4471	·73
·28	·5621	543·5592	·0065	·4379	·72
·29	·5714	561·7116	·0063	·4286	·71
8·30	8·5806	580·5283	0·0061	1·4194	1·70
·31	·5899	600·0353	·0060	·4101	·69
·32	·5992	620·2599	·0058	·4008	·68
·33	·6084	641·2302	·0056	·3916	·67
·34	·6177	662·9758	·0054	·3823	·66
8·35	8·6270	685·5274	0·0053	1·3730	1·65
·36	·6363	708·9171	·0051	·3637	·64
·37	·6456	733·1780	·0050	·3544	·63
·38	·6548	758·3451	·0048	·3452	·62
·39	·6641	784·4545	·0047	·3359	·61
8·40	8·6734	811·5439	0·0045	1·3266	1·60
·41	·6827	839·6528	·0044	·3173	·59
·42	·6920	868·8222	·0042	·3080	·58
·43	·7013	899·0948	·0041	·2987	·57
·44	·7106	930·5153	·0040	·2894	·56
8·45	8·7200	963·1301	0·0038	1·2800	1·55
·46	·7293	996·9878	·0037	·2707	·54
·47	·7386	1032·1389	·0036	·2614	·53
·48	·7479	1068·6362	·0035	·2521	·52
·49	·7572	1106·5347	·0034	·2428	·51
8·50	8·7666	1145·8919	0·0033	1·2334	1·50
		$1/Z$ Range	Z^2/PQ Weighting coefficient	$Y-P/Z$ Minimum working probit	Y Expected probit

Expected probit Y	Maximum working probit $Y+Q/Z$	Range $1/Z$	Weighting coefficient Z^2/PQ		
8·50	8·7666	1145·8919	0·0033	1·2334	1·50
·51	·7759	1186·7675	·0032	·2241	·49
·52	·7852	1229·2242	·0031	·2148	·48
·53	·7946	1273·3271	·0030	·2054	·47
·54	·8039	1319·1443	·0029	·1961	·46
8·55	8·8133	1366·7467	0·0028	1·1867	1·45
·56	·8226	1416·2085	·0027	·1774	·44
·57	·8320	1467·6071	·0026	·1680	·43
·58	·8413	1521·0232	·0025	·1587	·42
·59	·8507	1576·5411	·0024	·1493	·41
8·60	8·8600	1634·2488	0·0024	1·1400	1·40
·61	·8694	1694·2383	·0023	·1306	·39
·62	·8788	1756·6055	·0022	·1212	·38
·63	·8881	1821·4507	·0021	·1119	·37
·64	·8975	1888·8785	·0021	·1025	·36
8·65	8·9069	1958·9983	0·0020	1·0931	1·35
·66	·9162	2031·9243	·0019	·0838	·34
·67	·9256	2107·7758	·0019	·0744	·33
·68	·9350	2186·6775	·0018	·0650	·32
·69	·9444	2268·7596	·0017	·0556	·31
8·70	8·9538	2354·1583	0·0017	1·0462	1·30
·71	·9632	2443·0158	·0016	·0368	·29
·72	·9726	2535·4807	·0016	·0274	·28
·73	·9820	2631·7085	·0015	·0180	·27
·74	·9914	2731·8615	·0015	·0086	·26
8·75	9·0008	2836·1096	0·0014	0·9992	1·25
·76	·0102	2944·6302	·0014	·9898	·24
·77	·0196	3057·6091	·0013	·9804	·23
·78	·0290	3175·2401	·0013	·9710	·22
·79	·0384	3297·7264	·0012	·9616	·21
8·80	9·0478	3425·2801	0·0012	0·9522	1·20
·81	·0572	3558·1233	·0011	·9428	·19
·82	·0667	3696·4883	·0011	·9333	·18
·83	·0761	3840·6179	·0011	·9239	·17
·84	·0855	3990·7662	·0010	·9145	·16
8·85	9·0949	4147·1994	0·0010	0·9051	1·15
·86	·1044	4310·1955	·0010	·8956	·14
·87	·1138	4480·0457	·0009	·8862	·13
·88	·1232	4657·0549	·0009	·8768	·12
·89	·1327	4841·5419	·0009	·8673	·11
8·90	9·1421	5033·8407	0·0008	0·8579	1·10
·91	·1516	5234·3007	·0008	·8484	·09
·92	·1610	5443·2878	·0008	·8390	·08
·93	·1704	5661·1851	·0007	·8296	·07
·94	·1799	5888·3938	·0007	·8201	·06
8·95	9·1894	6125·3338	0·0007	0·8106	1·05
·96	·1988	6372·4452	·0007	·8012	·04
·97	·2083	6630·1886	·0006	·7917	·03
·98	·2177	6899·0468	·0006	·7823	·02
·99	·2272	7179·5252	·0006	·7728	·01
9·00	9·2367	7472·1536	0·0006	0·7633	1·00
		$1/Z$ Range	Z^2/PQ Weighting coefficient	$Y-P/Z$ Minimum working probit	Y Expected probit

The proportion killed is $p = 119/281 = 0{\cdot}4235$.

For $Y = 4{\cdot}61$, the Table shows the minimum working probit and range as

$$Y_{\text{min.}} = 3{\cdot}6680,$$

and $$1/Z = 2{\cdot}7047.$$

Hence the working probit is $$Y = 3{\cdot}6680 + 2{\cdot}7047p$$
$$= 4{\cdot}8134.$$

Alternatively, if survivors instead of deaths have been recorded, calculation may proceed from $$Y_{\text{max.}} = 6{\cdot}3727,$$

giving $$y = 6{\cdot}3727 - 2{\cdot}7047 \times 0{\cdot}5765$$
$$= 4{\cdot}8134.$$

The Table shows $$w = 0{\cdot}6023,$$

so that the weight for the observation is

$$nw = 281 \times 0{\cdot}6023$$
$$= 169{\cdot}2.$$

REFERENCES

BLISS, C. I. (1935*a*). The calculation of the dosage-mortality curve. *Ann. Appl. Biol.* **22**, 134–67.

BLISS, C. I. (1935*b*). The comparison of dosage mortality data. *Ann. Appl. Biol.* **22**, 307–33.

BLISS, C. I. (1938). The determination of dosage-mortality curves from small numbers. *Quart. J. Pharm.* **11**, 192–216.

FINNEY, D. J. (1947*a*). *Probit Analysis: A Statistical Treatment of the Sigmoid Response Curve.* London: Cambridge University Press.

FINNEY, D. J. (1947*b*). The principles of biological assay. *J. Roy. Statist. Soc. Suppl.* **9**, 46–91.

FISHER, R. A. (1935). Appendix to BLISS, C. I.: The case of zero survivors. *Ann. Appl. Biol.* **22**, 164–5.

FISHER, R. A. & YATES, F. (1943). *Statistical Tables for Biological, Agricultural and Medical Research* (2nd ed.). Edinburgh: Oliver and Boyd.

GADDUM, J. H. (1933). Reports on biological standards. III. Methods of biological assay depending on a quantal response. *Spec. Rep. Ser. Med. Res. Coun., Lond.*, no. 183.

MISCELLANEA

A note on the χ^2 smooth test

By H. L. SEAL

The now-classical test of goodness of fit of a theoretical frequency distribution to a set of observations consists of the calculation of $\sum_{j=1}^{n} (\delta m_j)^2/m_j$ (where m_j is the expectation of the jth group of observations and δm_j the deviation of the actual number of observations in this group from its expectation) and reference to tables of χ^2 with $n-1$ degrees of freedom. This is under the assumption that the only restraint on the theoretical frequency curve is that $\sum_{j=1}^{n} \delta m_j = 0$. In her recent contribution to this *Journal*, F. N. David (1947) shows by a geometrical method that, under these circumstances, a test of the significance of sequences of like signs in $\delta m_j (j = 1, 2, \ldots, n)$ is, for practical purposes, stochastically independent of the χ^2-test. The restriction of the arguments to cover only the case of a single linear restraint is, however, unnecessary.

It is possible, in this connexion, to state a useful general theorem: *If $x_j (j = 1, 2, \ldots, n)$ are n random variables normally distributed about zero mean with unit variance, these variables being connected by means of k linear relations, the probability distribution of $q^2 = \sum_{j=1}^{n} x_j^2$ remains unaltered if we select only those samples in which the signs of x_j follow a specified pattern.*

The proof is immediate since

$$p\{x_1, x_2, \ldots, x_n\} = \frac{1}{2^{\frac{1}{2}(n-k)-1}\,\Gamma[\frac{1}{2}(n-k)]}\, e^{-\frac{1}{2}q^2} q^{n-k-1}\, F(u_1, u_2, \ldots, u_{n-1}),$$

where $x_j = q u_j \quad j = 1, 2, \ldots, n, \qquad \sum_{j=1}^{n} u_j^2 = 1$

and the $u_j (j = 1, 2, \ldots, n-1)$ are connected by k further relations (cp. Hald & Rasch, 1943). Thus, changes in the signs of the x's are reflected by changes only in the signs of the u's, the joint distribution of which is independent of that of q.

Naturally this theorem does not apply directly to the case considered by David since the χ^2-test is there only an approximation to a set of terms of a multinomial expansion. It is, moreover, easily seen that if the squares of n independent random variables, with the same arbitrary skew distribution law, are added, the probability distribution of the resulting sum will necessarily depend on the signs of these variables. However, the practical use of the χ^2-test as an approximation implies 'near independence' of the test for sequences of signs even when k parameters (instead of one) have been fitted and have reduced the degrees of freedom to $n-k$.

A further point may be made. David's method of combining the χ^2-test of goodness of fit with the sequence test envisages a frequency distribution $m_j (j = 1, 2, \ldots, n)$ where n is relatively small: in fact her tables for the application of the combined test are calculated for $n = 5(1)14$. There is, however, a closely analogous case where n may assume a value between about 30 and 75, namely in testing the efficacy of the graduation of a mortality table. The writer (1943) suggested that in this case a good test would consist of the application of a χ^2-test of goodness of fit with $n-k$ degrees of freedom and a sequence test in the form of a 2×2 contingency table and associated χ^2 value, as indicated by Stevens (1939) himself when he suggested this latter test. In answer to an inquiry, the writer said he was 'not sure of the complete independence of the tests' but in view of the preceding it may be said that the second χ^2 (one degree of freedom) is very nearly independent of the χ^2 value in the main test, so that the two can be conveniently combined by addition into one value with $n-k+1$ degrees of freedom and a useful probability judgement of 'smooth fit' obtained.

REFERENCES

David, F. N. (1947). A χ^2 'smooth' test for goodness of fit. *Biometrika*, **34**, 299.

Hald, A. & Rasch, G. (1943). Nogle Anvendelser af Transformationsmetoden i den normale Fordelings Teori. *Festskrift til Prof. J. F. Steffensen.* Copenhagen.

Seal, H. L. (1943). Tests of a mortality table graduation. *J. Inst. Actu.* **71**, 5.

Stevens, W. L. (1939). Distribution of groups in a sequence of alternatives. *Ann. Eugen., Lond.*, **9**, 10.

Rank correlation and product-moment correlation

By P. A. P. MORAN, *Institute of Statistics, Oxford University*

The sampling distribution of Spearman's coefficient of rank correlation, ρ_s, has been thoroughly studied in the case where every permutation of the ranks of one variate relative to another is equiprobable and more recently Höffding (1948) has shown that it tends to normality for large samples whatever the parent population. This note deals with the distribution, when the parent is normal, for any size of sample.

Let $(x_1, y_1) \ldots (x_n, y_n)$ be a sample of n pairs of values from a bivariate normal population with correlation coefficient ρ. To obtain ρ_s we replace $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ by their respective ranks and calculate the product moment correlation coefficient of the ranks as if they were variate values. We have then to discuss the distribution of ρ_s when ρ is known.

Consider first the expected value of ρ_s. Let $p(x_i), p(y_i)$ be the ranks of x_i and y_i. Write

$$H(t) = 0 \quad \text{for} \quad t \leqslant 0$$
$$= 1 \qquad t > 0.$$

Then
$$p(x_i) - 1 = \sum_{j=1}^{n} H(x_i - x_j). \tag{1}$$

ρ_s will be the correlation coefficient of the numbers $\{p(x_i) - 1\}$ and $\{p(y_i) - 1\}$. If we had defined $H(t)$ in such a way that $H(0) = 1$ we would have obtained $p(x_i)$ on the left-hand side of equation (1) but this greatly complicates later calculations. Now write

$$S = \sum_{i=1}^{n} (p(x_i) - 1)(p(y_i) - 1).$$

It is easy to verify that
$$\sum_{i=1}^{n} \{p(x_i) - 1\} = \tfrac{1}{2}n(n-1),$$

and
$$\sum_{i=1}^{n} \{p(x_i) - \overline{p}\}^2 = \tfrac{1}{12}n(n^2-1),$$

where $\overline{p}$ is the mean of the ranks.

It follows that
$$\rho_s = \frac{s - \frac{1}{4}n(n-1)^2}{\frac{1}{12}n(n^2-1)},$$

and to find $E(\rho_s)$ it is enough to find $E(s)$. Now

$$s = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} H(x_i - x_j) H(y_i - y_k).$$

The terms in this expansion for which $i = j$ or $i = k$ are zero. There remain only two cases to consider.

Case I. i, j, k all distinct. Then $x_i - x_j, y_i - y_k$ are distributed in a bivariate normal distribution with correlation coefficient equal to $\frac{1}{2}\rho$. $E\{H(x_i - x_j) H(y_i - y_k)\}$ will therefore be the chance that both these quantities are positive, that is, the integral of the probability density over the positive quadrant. This is known (Sheppard, 1898) to be equal to $\frac{1}{4}\{1 - \pi^{-1}\cos^{-1}\frac{1}{2}\rho\}$. There are clearly $n(n-1)(n-2)$ such terms.

Case II. $i \neq j = k$. Then $x_i - x_j$ and $y_i - y_k$ have a correlation coefficient ρ and the expectation of each term is $\frac{1}{4}\{1 - \pi^{-1}\cos^{-1}\rho\}$. The number of such terms is $n(n-1)$.

It follows that

$$E(s) = \tfrac{1}{4}n(n-1)(n-2)(1 - \pi^{-1}\cos^{-1}\tfrac{1}{2}\rho) + \tfrac{1}{4}n(n-1)(1 - \pi^{-1}\cos^{-1}\rho),$$

and so
$$E(\rho_s) = \frac{12}{n(n^2-1)}\{\tfrac{1}{4}n(n-1)(n-2)(1 - \pi^{-1}\cos^{-1}\tfrac{1}{2}\rho) + \tfrac{1}{4}n(n-1)(1 - \pi^{-1}\cos^{-1}\rho) - \tfrac{1}{4}n(n-1)^2\}$$
$$= \frac{6}{\pi}\left\{\frac{n-2}{n+1}\sin^{-1}\tfrac{1}{2}\rho + \frac{1}{n+1}\sin^{-1}\rho\right\}. \tag{2}$$

For $\rho = 0, 1$ this equals zero or unity as we would expect. Equation (2) may be compared with K. Pearson's approximate formula (1907) for turning rank correlation coefficients into product moment correlation coefficients, which he derived by using the correlation of grades. This is

$$\rho_s = \frac{6}{\pi}\sin^{-1}\tfrac{1}{2}\rho.$$

Equation (2) tends to this as n increases, showing that ρ_s is not only a biased estimator of ρ, but also an inconsistent one. This, of course, does not prevent the use of ρ_s in estimating ρ.

The table which follows is a table of $E(\rho_s)$ as given by (2) for $n = 5, 10, 20$ and ∞, $\rho = 0(0{\cdot}1)\,0{\cdot}9$. The lower part of the table, for comparison, shows the expected values of r, the sample product-moment correlation coefficient, for the same values. These were taken from the tables in Soper, etc. (1917). Interpolation for $E(\rho_s)$ is linear in $(n+1)^{-1}$ and nearly linear in n^{-1}.

$\rho =$	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9
$E(\rho_s)$ $n = 5$	0·0797	0·1597	0·2408	0·3233	0·4080	0·4958	0·5883	0·6881	0·8022
$n = 10$	0·0869	0·1741	0·2620	0·3511	0·4419	0·5349	0·6313	0·7326	0·8428
$n = 20$	0·0910	0·1823	0·2742	0·3671	0·4613	0·5573	0·6559	0·7580	0·8659
$n = \infty$	0·0955	0·1913	0·2876	0·3849	0·4826	0·5819	0·6829	0·7859	0·8915
$E(r)$ $n = 5$	0·0884	0·1773	0·2671	0·3584	0·4517	0·5480	0·6482	0·7541	0·8687
$n = 10$	0·0946	0·1896	0·2850	0·3813	0·4787	0·5776	0·6785	0·7819	0·8887
$n = 20$	0·0974	0·1950	0·2928	0·3911	0·4900	0·5896	0·6902	0·7919	0·8951

Before considering the calculation of var (ρ_s), which we can find from var (s) and so from $E(s^2)$, we must consider the problem of evaluating the total probability in the positive part of a quadrivariate normal distribution.

Suppose that x_1, x_2, x_3, x_4 are distributed in a quadrivariate normal distribution with correlation matrix

$$\begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ & 1 & \rho_{23} & \rho_{24} \\ & & 1 & \rho_{34} \\ & & & 1 \end{pmatrix}.$$

We write

$$T\begin{pmatrix} \rho_{12} & \rho_{13} & \rho_{14} \\ & \rho_{23} & \rho_{24} \\ & & \rho_{34} \end{pmatrix} = pr\{x_1>0, x_2>0, x_3>0, x_4>0\}.$$

This will be independent of the variances of the x's which we suppose all equal to unity. Then the characteristic function of the distribution is

$$\phi(t_1,t_2,t_3,t_4) = \exp\{-\tfrac{1}{2}t_1^2-\tfrac{1}{2}t_2^2-\tfrac{1}{2}t_3^2-\tfrac{1}{2}t_4^2-\rho_{12}t_1t_2-\rho_{13}t_1t_3-\rho_{14}t_1t_4-\rho_{23}t_2t_3-\rho_{24}t_2t_4-\rho_{34}t_3t_4\},$$

and we then have

$$T\begin{pmatrix} \rho_{12} & \rho_{13} & \rho_{14} \\ & \rho_{23} & \rho_{24} \\ & & \rho_{34} \end{pmatrix}$$

$$= \frac{1}{16\pi^4}\int_0^\infty\int_0^\infty\int_0^\infty\int_0^\infty dx_1dx_2dx_3dx_4\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi(t_1,t_2,t_3,t_4)\exp\{-i\sum_1^4 t_ix_i\}\,dt_1dt_2dt_3dt_4.$$

Now

$$\phi(t_1,t_2,t_3,t_4)$$

$$= \exp\left\{-\tfrac{1}{2}\sum_1^4 t_i^2\right\}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{p=0}^{\infty}\sum_{q=0}^{\infty}\sum_{r=0}^{\infty}\frac{(-1)^{l+m+n+p+q+r}\rho_{12}^l\rho_{13}^m\rho_{14}^n\rho_{23}^p\rho_{24}^q\rho_{34}^r}{l!\,m!\,n!\,p!\,q!\,r!}$$

$$\times t_1^{l+m+n}t_2^{l+p+q}t_3^{m+p+r}t_4^{n+q+r},$$

and we can therefore write

$$T\begin{pmatrix} \rho_{12} & \rho_{13} & \rho_{14} \\ & \rho_{23} & \rho_{24} \\ & & \rho_{34} \end{pmatrix} = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{p=0}^{\infty}\sum_{q=0}^{\infty}\sum_{r=0}^{\infty}A_{lmnpqr}\rho_{12}^l\rho_{13}^m\rho_{14}^n\rho_{23}^p\rho_{24}^q\rho_{34}^r,$$

where

$$A_{lmnpqr} = \frac{(-1)^{l+m+n+p+q+r}G_{l+m+n}G_{l+p+q}G_{m+p+r}G_{n+q+r}}{l!\,m!\,n!\,p!\,q!\,r!},$$

and

$$G_s = \frac{1}{2\pi}\int_0^\infty dx\int_{-\infty}^{\infty}t^s\exp\{-\tfrac{1}{2}t^2-itx\}\,dt.$$

Now
$$\frac{1}{2\pi}\int_{-\infty}^{\infty} t^s \exp\{-\tfrac{1}{2}t^2 - itx\}\,dt = \frac{1}{2\pi(-i)^s}\left(\frac{d}{dx}\right)^s \int_{-\infty}^{\infty} \exp\{-\tfrac{1}{2}t^2 - itx\}\,dt$$
$$= \frac{1}{(-i)^s}\left(\frac{d}{dx}\right)^s \frac{1}{(2\pi)^{\frac{1}{2}}} e^{-\frac{1}{2}x^2}.$$

From this it follows that for $s = 0$, $G_s = \frac{1}{2}$. Consider $s > 0$. Then

$$G_s = \int_0^{\infty} \frac{1}{(-i)^s(2\pi)^{\frac{1}{2}}}\left(\frac{d}{dx}\right)^s e^{-\frac{1}{2}x^2}\,dx.$$
$$= \frac{-1}{(-i)^s(2\pi)^{\frac{1}{2}}}\left[\left(\frac{d}{dx}\right)^{s-1} e^{-\frac{1}{2}x^2}\right]_{x=0}$$
$$= 0 \qquad \text{when } s \text{ is even}$$
$$= \frac{(2m)!}{i(2\pi)^{\frac{1}{2}}\,2^m m!} \quad \text{when } s \text{ is odd and equal to } 2m+1, \text{ where } m = 0, 1, 2, \ldots.$$

Our final result is therefore

$$T\begin{pmatrix} \rho_{12} & \rho_{13} & \rho_{14} \\ & \rho_{23} & \rho_{24} \\ & & \rho_{34} \end{pmatrix} = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{p=0}^{\infty}\sum_{q=0}^{\infty}\sum_{r=0}^{\infty} \frac{(-1)^{l+m+n+p+q+r} G_{l+m+n} G_{l+p+q} G_{m+p+r} G_{n+q+r}}{l!\,m!\,n!\,p!\,q!\,r!} \rho_{12}^{l}\rho_{13}^{m}\rho_{14}^{n}\rho_{23}^{p}\rho_{24}^{q}\rho_{34}^{r}$$
$$= \frac{1}{16} + \frac{1}{8\pi}(\rho_{12}+\rho_{13}+\rho_{14}+\rho_{23}+\rho_{24}+\rho_{34})$$
$$+ \frac{1}{4\pi^2}(\rho_{12}\rho_{34}+\rho_{13}\rho_{24}+\rho_{14}\rho_{23}) + \ldots, \tag{3}$$

but not all the terms are positive. For example, the coefficient of

$$\rho_{12}\rho_{13}\rho_{14} (l = m = n = 1, p = q = r = 0) \quad \text{is} \quad -\frac{1}{4\pi^2}.$$

Expressions similar to the above have also been given by M. G. Kendall (1941, 1945) and have also been given in the lecture courses of Prof. A. C. Aitken as is stated in Kendall (1941). Kendall's version of the formula for four variables needs to be interpreted in the light of an earlier comment that the leading term is conventionally determined and in any case omits a power of -1.

We now consider the problem of finding $\operatorname{var}(\rho_s)$. Since

$$\operatorname{var}(\rho_s) = \frac{144}{n^2(n^2-1)^2}\operatorname{var}(s)$$

it will be sufficient to find $\operatorname{var}(s) = E(s^2) - [E(s)]^2$. Now

$$s^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{p=1}^{n}\sum_{q=1}^{n}\sum_{r=1}^{n} H(x_i - x_j)\,H(y_i - y_k)\,H(x_p - x_q)\,H(y_p - y_r). \tag{4}$$

Terms in this expansion for which $i = j$, $i = k$, $p = q$ or $p = r$ are zero. If we write

$$X_1 = x_i - x_j, \quad X_2 = y_i - y_k, \quad X_3 = x_p - x_q, \quad X_4 = y_p - y_r,$$

the quantities X_1, X_2, X_3, X_4 will be distributed in a quadrivariate normal distribution and the chance that they are all positive will be

$$T\begin{pmatrix} \frac{1}{2}\rho(1+\delta_{jk}) & \frac{1}{2}(\delta_{ip}-\delta_{iq}-\delta_{jp}+\delta_{jq}) & \frac{1}{2}\rho(\delta_{ip}-\delta_{ir}-\delta_{jp}+\delta_{jr}) \\ & \frac{1}{2}\rho(\delta_{ip}-\delta_{iq}-\delta_{kp}+\delta_{kq}) & \frac{1}{2}(\delta_{ip}-\delta_{ir}-\delta_{kp}+\delta_{kr}) \\ & & \frac{1}{2}\rho(1+\delta_{pr}) \end{pmatrix}, \tag{5}$$

where δ_{ij} is the Kronecker δ equal to unity if $i = j$ and zero otherwise. Using (3) we can now evaluate

$$E\{H(x_i - x_j)\,H(y_i - y_k)\,H(x_p - x_q)\,H(y_p - y_r)\} = pr\{X_1 > 0, X_2 > 0, X_3 > 0, X_4 > 0\},$$

and we must calculate (5), using (3), for each type of set of suffixes in the terms of (4). The number of times each such type of term occurs in (4) will be $n(n-1)\ldots(n-s+1)$, where s is the number of distinct suffixes. From a theoretical point of view this solves the problem but in practice, since there are a fairly large number of different terms which arise from the various ways of identifying suffixes in (4), it is probably only barely possible to calculate $\operatorname{var}(\rho_s)$ for any given value of ρ. It is certainly not practical as a routine.

The above investigation may be compared with the similar investigations on the sampling distribution of Kendall's τ (Greiner, 1909; Esscher, 1924; Kendall, 1948) and the relative simplicity of the latter, especially in the formulae for the variance, brings out once again the superiority of τ as a measure of rank correlation.

REFERENCES

Esscher, F. (1924). On a method of determining correlation from the ranks of variables. *Skand. AktuarTidskr.* **7**, 201.

Greiner, R. (1909). Über das Fehlersystem der Kollektiv-masslehre. *Z. Math. Phys.* **57**, 225.

Höffding, W. (1948). A class of statistics with asymptotically normal distributions. *Ann. Math. Statist.* (In the Press.)

Kendall, M. G. (1941). Proof of relations connected with the tetrachoric series and its generalization. *Biometrika*, **32**, 196.

Kendall, M. G. (1943). *The Advanced Theory of Statistics*, vol. 1 (3rd ed. 1947). Charles Griffin and Co.

Kendall, M. G. (1945). On the analysis of oscillatory time-series. *J.R. Statist. Soc.* **108**, 93.

Kendall, M. G. (1948). *Rank Correlation Methods.* London: Charles Griffin & Co. (In the Press.)

Pearson, K. (1907). Mathematical contributions to the theory of evolution. XVI. On further methods of determining correlation. *Drap. Co. Mem. biom. Ser.* Cambridge University Press.

Sheppard, W. F. (1898). On the application of the theory of error to cases of normal distributions and normal correlations. *Philos. Trans.* A, **192**, 101.

Soper, H. E. and others (1917). On the distribution of the correlation coefficient in small samples. *Biometrika*, **11**, 328.

Tests of significance in the variate difference method

By N. L. JOHNSON

1. The variate difference method has been discussed by many writers, and in particular by Yule (1921), Anderson (1923, 1926, 1927) and Tintner (1940). It is a method of isolating the random part of certain types of time series and is briefly described in § 2 below.

2. Let $u_1, u_2, \ldots, u_n$ be successive observations at equal intervals of time, forming a time series. Suppose that it is reasonable to assume that

$$u_t = f_t + z_t, \tag{1}$$

where (i) f_t is a 'smooth trend' such that for some integer K (and hence for all larger integers)

$$\Delta^K f_t = 0;$$

(ii) z_t is a 'random residual' with expected value zero and standard deviation σ, independent of t;
(iii) $z_1, z_2, \ldots, z_n$ are mutually independent.

Under these conditions, if the series (u_t) be differenced K times, the f_t terms will be eliminated and, in fact,

$$\Delta^K u_t = \Delta^K z_t. \tag{2}$$

Furthermore $\Delta^{K+1} u_t = \Delta^{K+1} z_t$, $\Delta^{K+2} u_t = \Delta^{K+2} z_t$ and so on.

Since these relations hold, we have, for $k \geqslant K$,

$$E(\Delta^k u_t) = 0, \tag{3·1}$$

$$E[\{\Delta^k u_t\}^2] = {}^{2k}C_k \sigma^2. \tag{3·2}$$

Thus, if $k \geqslant K$

$$S_k = [(n-k)\,{}^{2k}C_k]^{-1} \sum_{t=1}^{n-k} \{\Delta^k u_t\}^2 \tag{4}$$

is an unbiased estimate of σ^2.

If $\Delta f_t = 0$, i.e. if the original series is random

$$S_0 = (n-1)^{-1} \sum_{t=1}^{n} (u_t - \bar{u})^2 \tag{5}$$

is an unbiased estimate of σ^2.

The sequence of statistics $S_0, S_1, S_2 \ldots$ is computed from the observed values $u_1, u_2, \ldots, u_n$; S_k being defined by (4) for $k \geqslant 1$ and by (5) for $k = 0$. This sequence may be used as an indicator of the order of difference at which the non-random elements are first eliminated. Generally the original series is not

random and so S_1 is usually considerably smaller than S_0. As k increases the ratio S_{k+1}/S_k should approach unity. The value of k at and after which S_{k+1}/S_k stays sufficiently close to unity is taken as the order of difference necessary to eliminate the non-random terms f_t.

3. In order to test precisely whether the ratio S_{k+1}/S_k is in fact 'sufficiently close' to unity it is necessary to assume a form of distribution for the z_t's. It is supposed that the assumption of a normal form of distribution for the z_t's will not give rise to serious error. Even with this assumption it is very difficult to obtain exact significance levels for the ratio S_{k+1}/S_k, owing to the existence of correlations of various degrees of intensity between the differences involved.

Tintner suggested a method of overcoming this drawback. He proposed that certain sets of the kth and $(k+1)$th differences should be selected in such a way that any selected difference of either order should be uncorrelated with any other selected difference of either order. The sets to be selected were of the type:

$$(\Delta^k u_r) \quad r = t, \quad t+(2k+3), \quad t+2(2k+3), \quad \ldots, \quad t+(j-1)(2k+3);$$

$$(\Delta^{k+1} u_s) \quad s = t+k+1, \quad t+k+1+(2k+3), \quad \ldots, \quad t+k+1+(j-1)(2k+3).$$

t may have any integral value from 1 to $(2k+3)$ inclusive, each value giving rise to a different selection; j has the largest value possible. In any one selection none of the quantities $\Delta^k u_r, \Delta^{k+1} u_s$ have a u_t in common, and their independence is thereby assured. If the non-random elements have been removed by taking kth differences, then $(\Delta^k u_r), (\Delta^{k+1} u_s)$ should be sequences of independent normal variables, each with expected value zero and with variances in the ratio

$$^{2k}C_k / {}^{2k+2}C_{k+1} = \tfrac{1}{4}(k+1)/(2k+1).$$

Tintner suggests using as test criterion

$$F = \frac{\sum_r \{\Delta^k u_r\}^2}{\sum_s \{\Delta^{k+1} u_s\}^2} \frac{2(2k+1)}{k+1} \quad \text{or} \quad z = \tfrac{1}{2}\log_e F. \tag{6}$$

Standard tables of F or z, entered with degrees of freedom j, j would then provide exact significance limits for the suggested criteria.

4. Although this method of selection gives a test which is exact, provided the assumptions upon which it is based are correct, it involves the sacrifice of a considerable proportion of the data available. Thus only about one out of each $(2k+3)$ members of the two difference columns under comparison is used explicitly in the test criteria. This is necessary if the correlation between *any* two selected differences is to be zero.

It is possible, however, to make use of somewhat more of the data available, still subjecting the results to a fairly simple exact test, analogous to that proposed by Tintner. Consider, in fact, the modified method of selection leading to the sequence of pairs of differences

$$(\Delta^k u_r, \Delta^{k+1} u_r) \quad r = t, \quad t+k+2, \quad t+2(k+2)\ldots, \quad t+(j'-1)(k+2). \tag{7}$$

Here t may be any integer from 1 to $(k+2)$ inclusive, and j' has its largest possible value.

If non-random elements have been eliminated in the kth differences then

(i) the correlation between corresponding differences $\Delta^k u_r$ and $\Delta^{k+1} u_r$ is $-[\frac{1}{2}(2k+1)/(k+1)]^{\frac{1}{2}}$;

(ii) any difference of one order is correlated only with the corresponding difference of the other order;

(iii) the expected values and ratio of variances of the sequences $(\Delta^k u_r), (\Delta^{k+1} u_r)$ are the same as in Tintner's method.

An exact test based on the selection (7) would utilize about twice as many differences as would Tintner's test for the same time series. For example, we have the following comparison between Tintner's method of selection and method (7) in the special case $t = 1, k = 2$.

Tintner's Method		Present Method (7)	
$\Delta^2 u_r$	$\Delta^3 u_s$	$\Delta^2 u_r$	$\Delta^3 u_r$
$\Delta^2 u_1 = u_3 - 2u_2 + u_1$	$\Delta^3 u_4 = u_7 - 3u_6 + 3u_5 - u_4$	$\Delta^2 u_1 = u_3 - 2u_2 + u_1$	$\Delta^3 u_1 = u_4 - 3u_3 + 3u_2 - u_1$
$\Delta^2 u_8 = u_{10} - 2u_9 + u_8$	$\Delta^3 u_{11} = u_{14} - 3u_{13} + 3u_{12} - u_{11}$	$\Delta^2 u_5 = u_7 - 2u_6 + u_5$	$\Delta^3 u_5 = u_8 - 3u_7 + 3u_6 - u_5$
etc.	etc.	etc.	etc.
Variance $6\sigma^2$	$20\sigma^2$	$6\sigma^2$	$20\sigma^2$
Correlation 0		$-\sqrt{(5/6)}$	

The gain in efficient use of the data would not, however, be as great as this would suggest, since the correlation is high between the differences already in Tintner's selection and the additional differences introduced in (7). Nevertheless, a certain amount of extra information would be used in a test based on (7). Such an exact test, analogous to Tintner's test, is derived in § 5.

5. In the case of Tintner's method of selection, the hypothesis which should be tested is that the expected values of each of the differences is zero. The ratio of the variances of the differences in the two sequences is known. The alternative hypotheses specify non-zero values of the expected values, the ratio of the variances remaining unchanged. It is not necessary that these non-zero expected values be constant for either of the sequences of differences involved. Tintner's criterion is, however, appropriate to the case where the alternative hypotheses specify different values for the ratio of the variances, the expected values remaining at zero (i.e. the hypothesis tested is that the variances are in a certain ratio, it being assumed that the expected values are zero). This system of hypotheses seems to be a reasonable approximation to the situation, as may be appreciated by regarding the differences $\Delta^k f_r, \Delta^{k+1} f_r$, as random variables.

A similar approach in the case of selection (7) leads to the conclusion that the hypothesis to be tested may be stated as follows: 'The variances of the two sequences of differences are in a certain ratio, and the correlation between corresponding differences has a certain value. It is assumed that all expected values are zero.' The alternative hypotheses specify other values for the ratio of the variances and the correlation, all expected values being supposed to remain at zero.

A test appropriate to this situation may be obtained by a slight modification of a result due to Hsu (1940), who used Neyman and Pearson's likelihood ratio method to derive a number of tests of hypotheses regarding two normally correlated variables. The hypothesis to be tested is, in fact, nearly the same as Hsu's hypothesis H_4 and the appropriate test criterion is similar to his criterion L_4. This similarity is not, however, apparent when the special symbols and values of the present problem are inserted, giving a test criterion

$$L = \frac{T_{k,k}\,T_{k+1,k+1} - T_{k,k+1}^2}{(2k+1)\,[T_{k,k} + \frac{1}{2}(k+1)\,(2k+1)^{-1}\,T_{k+1,k+1} + T_{k,k+1}]^2}, \tag{8}$$

where $T_{p,q} = \sum_r \Delta^p u_r \Delta^q u_r$.

L must lie between 0 and 1. Low values of L are regarded as significant of departure from the hypothesis tested, indicating that the 'non-random' terms have not been eliminated in the kth differences. If non-random variations are in fact absent from the kth differences, the probability density function of L is

$$p(L) = \tfrac{1}{2}(j'-1)\,L^{\frac{1}{2}(j'-3)} \quad (0 \leqslant L \leqslant 1). \tag{9}$$

It will be recalled that j' is the number of differences in each of the sequences $(\Delta^k u_r), (\Delta^{k+1} u_r)$. From (9) it follows that L_α, the significance limit corresponding to a probability α of rejecting the hypothesis when it is valid, satisfies the equation $L_\alpha^{\frac{1}{2}(j'-1)} = \alpha$,

i.e.
$$\log_{10} L_\alpha = (2 \log_{10} \alpha)/(j'-1). \tag{10}$$

If $\log_{10} L$ be taken as test criterion, the significance limits may be calculated very easily by means of (10). We note that the 5 % limit for $\log_{10} L$ is $-2{\cdot}6/(j'-1)$, and the 1 % limit is $-4{\cdot}0/(j'-1)$.

6. We shall now apply selection (7) to the series of American wheat-flour prices for the years 1890–1937, which is analysed by Tintner. There are forty-eight observations in this series. Taking $t = 1$, and comparing the first and second order differences (i.e. $k = 1$) we have the sequences

$$\Delta u_1, \quad \Delta u_4, \quad \Delta u_7, \quad \ldots, \quad \Delta u_{46};$$
$$\Delta^2 u_1, \quad \Delta^2 u_4, \quad \Delta^2 u_7, \quad \ldots, \quad \Delta^2 u_{46};$$

so that $j' = 16$. (It may be noted that the sequences of differences in Tintner's selection 1-A each contain 9 members.) The sums of squares and products are

$$T_{1,1} = 31{\cdot}259830, \quad T_{1,2} = -23{\cdot}842300, \quad T_{2,2} = 29{\cdot}137173,$$

whence
$$L = 0{\cdot}3889,$$

and
$$\log_{10} L = -0{\cdot}4102.$$

For $j' = 16$, the 1 % limit for $\log_{10} L$ is $-0{\cdot}27$ and the 5 % limit is $-0{\cdot}17$. The calculated value of $\log_{10} L$ is less than the 1 % limit, and it seems unlikely that the non-random terms have been eliminated in the first differences.

Now taking $k = 2$ (and $t = 1$ as before) we have the sequences of differences

$$\Delta^2 u_1, \quad \Delta^2 u_5, \quad \Delta^2 u_9, \quad \ldots, \quad \Delta^2 u_{45}; \qquad \Delta^3 u_1, \quad \Delta^3 u_5, \quad \Delta^3 u_9, \quad \ldots, \quad \Delta^3 u_{45};$$

so that in this case $j' = 12$. The sums of squares and products are

$$T_{2,2} = 12{\cdot}030127; \quad T_{2,3} = -10{\cdot}330714; \quad T_{3,3} = 53{\cdot}809942$$

whence
$$L = 0{\cdot}3396,$$

and
$$\log_{10} L = -0{\cdot}4690.$$

For $j' = 12$, the 1 % limit for $\log_{10} L$ is $-0{\cdot}36$ and the 5 % limit is $-0{\cdot}24$. Again it seems unlikely that non-random terms have been eliminated.

Comparing third and fourth order differences ($k = 3$) we have, using the selection corresponding to $t = 1$

$$T_{3,3} = 102{\cdot}200898, \quad T_{3,4} = -188{\cdot}615311, \quad T_{4,4} = 358{\cdot}329689,$$

giving $L = 0{\cdot}5862$ and $\log_{10} L = -0{\cdot}2320$. The appropriate 1 % limit ($j' = 9$) is $-0{\cdot}50$ and the 5 % limit is $-0{\cdot}33$. This test supports the hypothesis that non-random terms have been eliminated in the third order differences. Before a final decision is reached, of course, further tests would be made, using other possible values of t, and dealing with higher orders of differences. The above calculations should, however, be sufficient to make clear the mode of application of the method of selection (7) and its associated test of significance.

7. It is not essential that the sequence of pairs of differences be chosen as in (7) above. Any sequence satisfying conditions (ii) and (iii) of § 4 may be used, and a corresponding criterion, similar to (8), obtained by inserting the appropriate terms in the $T_{p,q}$'s and using the correct value of the correlation coefficient between corresponding differences. The sequence

$$(\Delta^k u_{r+1}, \Delta^k u_r) \quad r = t, \quad t+k+2, \quad t+2(k+2), \quad \ldots, \quad t+(j'-1)(k+2)$$

is very similar to (7) and leads, in fact, to identically the same criterion.

Other alternative methods of selection may aim at reducing the correlation between corresponding differences. For example, in the sequence

$$(\Delta^k u_r, \Delta^{k+1} u_{r+p}) \, (p \leqslant k) \quad r = t, \quad t+k+2+p, \quad t+2(k+2+p), \quad \ldots,$$

this correlation is

$$(-1)^{p+1} {}^{2k+1}C_{p+k+1} / ({}^{2k}C_k \, {}^{2(k+1)}C_{k+1})^{\frac{1}{2}} = (-1)^{p+1} \frac{k(k-1)\ldots(k-p+1)}{(k+2)(k+3)\ldots(k+p+1)} \left(\frac{2k+1}{2k+2}\right)^{\frac{1}{2}}.$$

Unfortunately the more the correlation is reduced the greater the proportion of data eliminated by this method of selection.

A further possible method, possessing the advantage of symmetry, could be based on the comparison of kth and $(k+2)$th order differences, using the sequence

$$(\Delta^k u_{r+1}, \Delta^{k+2} u_r) \quad r = t, \quad t+k+3, \quad t+2(k+3), \ldots,$$

in which the correlation between corresponding differences is

$$-[(2k+1)(k+2)/(2k+3)(k+1)]^{\frac{1}{2}}.$$

REFERENCES

Anderson, O. (1923). *Biometrika*, **15**, 134.

Anderson, O. (1926). *Biometrika*, **18**, 293.

Anderson, O. (1927). *Biometrika*, **19**, 53.

Hsu, C. T. (1940). *Ann. Math. Statist.* **11**, 410.

Tintner, G. (1940). *The Variate Difference Method.* Cowles Commission, Principia Press, Bloomington, Ind.

Yule, G. U. (1921). *J. Roy. Statist. Soc.* **84**, 497.

REVIEW

The Advanced Theory of Statistics. Vol. II, pp. 1–521. By M. G. KENDALL, M.A. London: Griffin and Co., Ltd. 1946. Price: 50*s*.

There is only a small measure of agreement among writers on the theory of statistics as to what should constitute an advanced course in the subject. There is no common view about either the nature of the problems which should be discussed or the methods by which they should be solved. Disagreement about the actual details of solutions of theoretical problems is, perhaps, only a passing phase and its importance may be exaggerated. The way in which the problems themselves are posed, however, reveals more fundamental cleavages of opinion which may not be so easily reconciled.

An important element which renders difficult the formulation of theoretical problems is the uncertain relation existing between the theory and the practice of the application of statistical methods, as soon as we advance beyond the most elementary stages. In discussing this relation it is convenient to distinguish between the theoretician or, as he is now categorized, the mathematical statistician, on the one hand, and the practical statistician, using statistical methods in some particular field of inquiry, on the other. This distinction may have some validity if it is not pushed too hard. It is true that at one extreme there is a body of workers whose interest in the theory of statistics is primarily as a source of mathematical problems. And, at the other extreme, there are experimentalists who apply statistical methods by rule of thumb, without much concern for the reasoning which is needed to justify them. But, although the modern tendencies in the organization of all the sciences may be such as to force workers into extreme positions and label them accordingly, it must be recognized that, in statistics, the theoretical and the practical investigator are still, fortunately, often one and the same person.

This fact undoubtedly makes for a closer integration of theoretical developments with practical usages, but not, perhaps, to the extent that one might imagine. It is in the nature of theoretical work—almost in its definition—that, whoever may be responsible for it, it should assert the right to an independent existence of its own and refuse to be tied down too closely by considerations of its ultimate usefulness. Even were statistical theory to be cultivated exclusively by writers who at the same time were outstandingly competent in dealing with the more commonplace, everyday, statistical investigations, it would still not escape the tendency to become a purely abstract discipline. The time spent on it would still have to be justified by secondary considerations, such as its aesthetic appeal, educational value and the like, as well as by its ability to provide novel statistical procedures of practical value in the experimental and social sciences.

The tendency of statistical theory towards exclusive concern with abstract generalizations is helped, of course, by those mathematicians who are interested in nothing else. There is, however, as one might expect, a strong reaction against it from statisticians who still maintain a wider outlook. This reaction expresses itself, sometimes explosively, in the form of exasperated criticism of this or that piece of theoretical work on the grounds, either of its irrelevance, or worse, of the misleading impressions which it may give of the true objects of practical statistical investigations. The criticism is not always fair or well directed. Its authors are seldom immune from the failings, real or imaginary, which they castigate in others. But it is sufficient to show that statisticians as a body, however pleased they may be that their subject is gaining increased academic recognition, are yet unwilling to sanction the appearance of two separate subjects—pure and applied statistics—which can develop without much reference one to the other. Practical interests are still able to make themselves felt even in academic developments. Further, since the practical interests of statisticians are so diverse, it is not surprising that a parallel difficulty is found in reaching agreement as to what is the proper subject-matter for the advanced theory of statistics.

Mr M. G. Kendall, in the notable work of which the second volume is here under review, does not attempt to confine the subject within any very strict limits, nor to press upon the reader any very decided views as to what should constitute its most important features. His method is rather to let the truth appear—if appear it can—by making an extensive and detailed survey of the whole range covered by modern contributions.

The first four chapters of this second volume are concerned with the fundamental problem of estimating population parameters from sample values and of assigning limits to such parameters on a probability basis. This has in the past produced a most controversial literature, revolving round the question of the applicability or otherwise of the theory of inverse probability and of the various alternatives to this theory which have been proposed. Mr Kendall devotes most attention to the approach which does not make explicit use of the concept of inverse probability. Some readers might

consider that he should have given more space to the alternative position which is founded on Bayes's Theorem and its corollaries. However, in a sense, Harold Jeffreys's excellent restatement of this position in his *Theory of Probability* renders a lengthy description of it unnecessary, unless, indeed, Mr Kendall's object had been to make a critical comparison and evaluation of the inverse and direct probability approaches. Mr Kendall is content to describe rather than to judge. One might, perhaps, criticize him in this section of the book for carrying his non-committal attitude too far in the face of the irreconcilable viewpoints which emerge from the discussion. He could, possibly, have pressed his own ideas upon us more strongly in certain passages, without departing from the high standard of fairness which he maintains throughout in representing the ideas of his statistical colleagues.

The problem of scientific inference discussed in these first four chapters, although a fertile source of argument, is, strangely enough, one which, in whatever way it is settled, does not appear to influence much the actual way in which scientific investigation is carried out in practice. The next five chapters are concerned with a miscellaneous range of questions, which, while not having the same attraction for the pure theorist, are more closely related to those statistical methods used most frequently in the everyday interpretation of experimental data. They include, for instance, some discussion of the theory of the analysis of variance and of the role which randomization plays in ensuring the applicability of this theory in practice. The attitude which one adopts towards the theoretical questions discussed in this part of the book can affect quite directly the choice between the alternative experimental procedures which may be at one's disposal in some specific inquiry.

The succeeding two chapters revert to a more generalized treatment following the same lines as the earlier ones. They are again concerned with the fundamental problem of scientific inference, although now from the restricted angle of significance testing. They furnish another illustration of the author's remarkable ability to reproduce the spirit as well as the substance of the original contributions covered by his survey. They again suffer, perhaps, from some lack of critical evaluation on the author's own part.

The next chapter deals with multivariate analysis, including discriminant functions. From the theoretical standpoint this is a straightforward development from the univariate and bivariate cases. Some new problems are raised, but much is simply generalization. Of the possible applications of these recent developments it is difficult to speak confidently. It is clear that in many fields multivariate analysis of the type here described can only proceed on the basis of many dubious assumptions as to the nature of possible connexions between the variates. Moreover, simple interpretation of the results of a multivariate analysis in terms which convey much to the layman is difficult. On the other hand, in certain cases where a very large number of variates are measured (e.g. in intelligence testing) some logical way of dealing with the results is needed, and one may in some instances be sure enough of one's assumptions to exploit this recent theoretical work.

The two final chapters deal with Time Series, a subject perhaps most studied by economists, but not one which has borne much fruit. Trade cycles have been sufficiently talked about, but there has been little analysis of data demonstrating their existence in an unequivocal fashion. Mr Kendall is the foremost representative of a school of thought which holds that the search for regular periodicities in economic data has been largely a waste of time, and that much more is to be hoped for by considering so-called autoregressive schemes, which allow for irregularities in the lengths of periods—albeit irregularities governed by some simple law. It is too soon yet to say whether the new conceptions will be much more successful than the old, but whatever success is obtained will be largely due to Mr Kendall's own efforts to clarify the subject.

Looking back after reading through the two volumes of Mr Kendall's ambitious enterprise, one is forced to the realization that here we have one of the most remarkable compilations that has ever been attempted by a single writer in any branch of science. If several writers had banded together to produce it one would still be impressed by the wide range of topics and viewpoints which it embraces. One might almost be pardoned, indeed, for thinking that 'M. G. Kendall' was a pseudonym standing for the collaboration of several persons. This would, at all events, be a pleasant theory to explain the many excellences of the work, particularly its freedom from personal animosities and the generosity of its references to the contributions of so many different writers. It would also explain the somewhat loose-knit nature of the work judged as a whole. Occasionally one feels the lack of a sufficiently strong common thread running through it and holding it together; such as one would perhaps obtain if the author were inclined to be more trenchant in his criticisms of what other people write and more self-assertive in putting forward his own contributions. But in reply to all possible criticisms, Mr Kendall has the final word, when, in his concluding section, he says: 'Much remains to be done; and this book will have served its purpose if the reader is left with the desire to do some of it himself.' Few authors could by implication have exposed so clearly just what it is that remains to be done.

B.L.W.

(*All Rights reserved*)

BIOMETRIKA. Vol. XXXV, Parts I and II

CONTENTS

	PAGE
James Fowler Tocher. (*With frontispiece*)	1–5
On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. By D. G. KENDALL	6–15
The studentized form of the extreme mean square test in the analysis of variance. By K. R. NAIR	16–31
The estimation of non-linear parameters by 'internal least squares'. By H. O. HARTLEY	32–45
The geometrical method in the theory of sampling. By DAVID FOG	46–54
Proofs of the distribution law of the second order moment statistics. By JOHN WISHART	55–57
Tests of significance in multivariate analysis. By C. RADHAKRISHNA RAO	58–79
Alternative systems in the analysis of variance. By N. L. JOHNSON	80–87
An examination and further development of a formula arising in the problem of comparing two mean values. By ALICE A. ASPIN	88–96
On the power function of the longest run as a test for randomness in a sequence of alternatives. By G. BATEMAN	97–112
Sur les courbes de fréquence de K. Pearson. PAR M. DUMAS	113–117
The distribution of the extreme deviate from the sample mean and its studentized form. By K. R. NAIR	118–144
The Fisher-Yates test of significance in 2×2 contingency tables. By D. J. FINNEY	145–156
The power function of the test for the difference between two proportions in a 2×2 table. By P. B. PATNAIK	157–175
The analysis of contingency tables with groupings based on quantitative characters. By F. YATES	176–181
The probability integral transformation when parameters are estimated from the sample. By F. N. DAVID and N. L. JOHNSON	182–190
A table for the calculation of working probits and weights in probit analysis. By D. J. FINNEY and W. L. STEVENS	191–201
MISCELLANEA	
A note on the χ^2 smooth test. By H. L. SEAL	202
Rank correlation and product-moment correlation. By P. A. P. MORAN	203–206
Tests of significance in the variate difference method. By N. L. JOHNSON	206–209
REVIEW	
The Advanced Theory of Statistics	210

A volume of *Biometrika* contains about 400 pages, with plates and tables, and it is hoped that in future this will be published annually in two half-yearly issues.

Papers for publication should either be sent to

PROFESSOR E. S. PEARSON, Department of Statistics, University College, London, W.C. 1,

or if more convenient to

DR JOHN WISHART, School of Agriculture, Cambridge.

or M. G. KENDALL, Esq., 9 Duke's Lodge, Holland Park, London, W. 11.

It is a condition of publication in *Biometrika* that the paper shall not already have been issued elsewhere, and will not be reprinted without leave of the Editors.

Contributors receive 25 copies of their papers free. Joint authors 15 copies each.

The subscription price, *payable in advance*, is Inland 45*s. net* per volume and Abroad 54*s. net* (including packing and postage). Owing to the scarcity of early volumes, the following rates must now be charged for complete sets. Vols. I—XXXIV, including XX^B: £130 in wrappers, not including postage. At present certain volumes are out of print, but steps are being taken to re-issue these as quickly as printing facilities permit. Recent volumes may still be obtained at the wrapper price; this is 6*s.* inland, including postage. Index to Vols. I to V, 2*s. net*. Index to Vols. I to XV, 5*s. net*. Cheques must be made payable to *Biometrika*, crossed "*a/c Biometrika Trust*" and sent to The Secretary, *Biometrika Office*, Department of Statistics, University College, London, W.C. 1, to whom all orders for series, single copies and offprints should be addressed. All foreign cheques must be drawn in sterling and on a Bank having a London Agency.

First printed in Great Britain at the University Press, Cambridge
Reprinted by offset-litho by Percy Lund, Humphries & Co., Ltd, Bradford